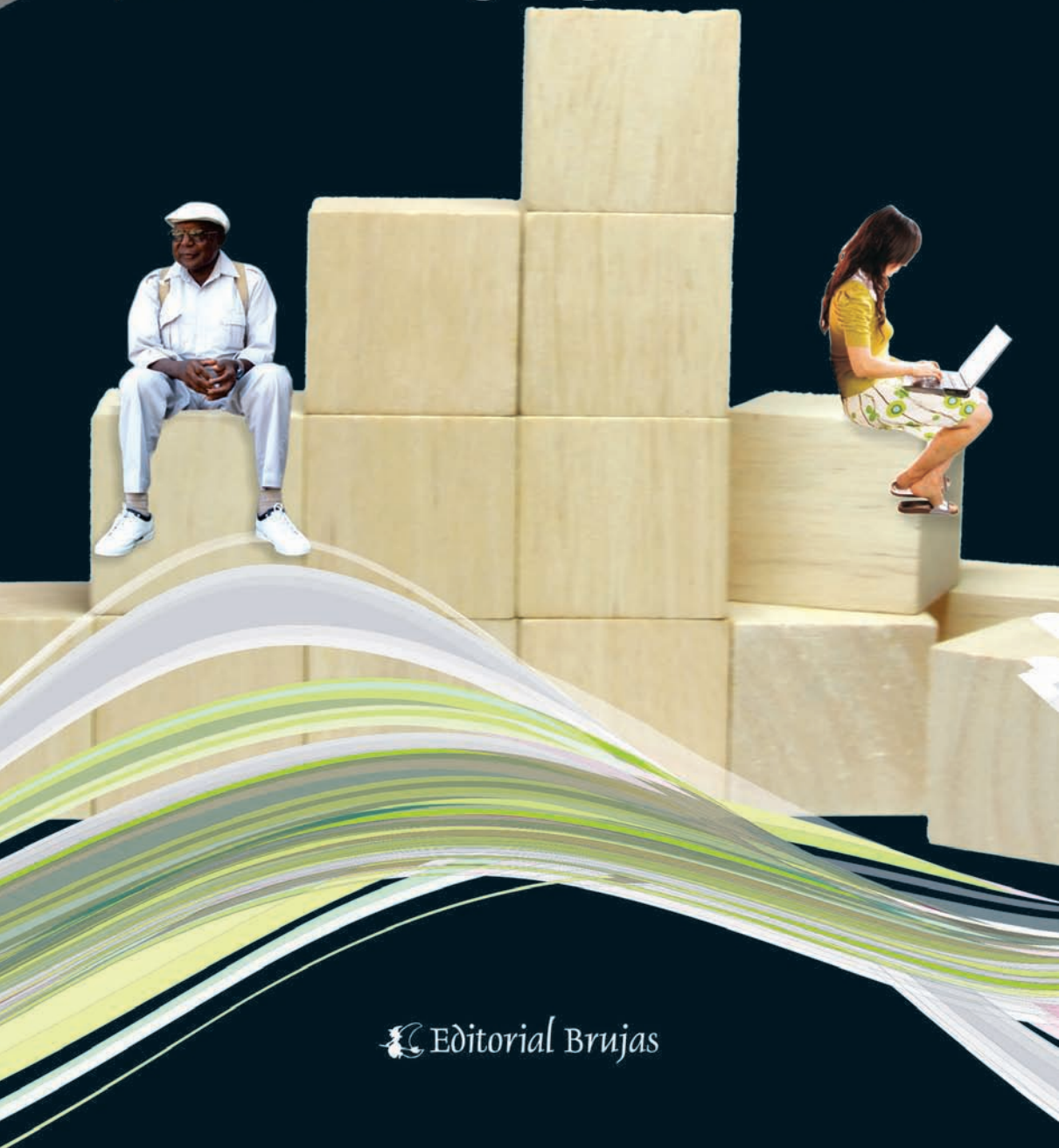
EDUARDO BOLOGNA

Estadística
para Psicología y Educación



Editorial Brujas

Eduardo Bologna

Estadística para Psicología y Educación

Editorial Brujas

La presente es la versión aumentada y corregida del texto Estadística en psicología. Ed. Brujas, Córdoba, 2010 ISBN 978-987-591-205-2

Título: Estadística para psicología y educación
Autor: Eduardo Bologna
Colaboradores
Faas, Ana Eugenia
González de Menne, María Cristina
Medrano, Leonardo
Morales, María Marta
Reyna, Cecilia
Romero, Waldino
Urrutia, Andrés

Bologna, Eduardo
Estadística para psicología y educación. - 1a ed. - Córdoba: Brujas, 2011.
454 p. ; 24x16 cm.

ISBN 978-987-591-249-6

1. Estadística. I. Título.
CDD 310

1° Edición.

Impreso en Argentina

ISBN:978-987-591-249-6

Queda hecho el depósito que marca la ley 11.723.

ENCUENTRO Grupo Editor

Editorial Brujas Miembros de la CÁMARA ARGENTINA DEL LIBRO

www.editorialbrujas.com.ar publicaciones@editorialbrujas.com.ar

Tel/fax: (0351) 4606044 / 4691616- Pasaje España 1485 Córdoba - Argentina.

"Un hombre de Neguá, en la costa de Colombia, pudo subir al alto cielo. A la vuelta contó. Dijo que había contemplado, desde allá arriba, la vida humana. Y dijo que somos un mar de fueguitos.
—El mundo es eso —reveló—, un montón de gente, un mar de fueguitos. Cada persona brilla con luz propia entre todas las demás. No hay dos fuegos iguales. Hay fuegos grandes y fuegos chicos y fuegos de todos los colores. Hay gente de fuego sereno, que ni se entera del viento, y gente de fuego loco, que llena el aire de chispas. Algunos fuegos, fuegos bobos, no alumbran ni queman; pero otros arden la vida con tantas ganas que no se puede mirarlos sin parpadear,
y quien se acerca, se enciende."
Eduardo Galeano, El libro de los abrazos

Agradecimientos

Desde el impulso a la idea de actualizar la bibliografía de Estadística en Psicología hasta la lectura detallada de los originales, la presencia del Profesor Livio Grasso ha sido determinante para la realización de este trabajo. Su interés por mejorar la forma de dictado y sus intentos para que los alumnos vean la utilidad de la estadística, así como que confíen en sus capacidades, fue un estímulo permanente en la redacción y corrección de este material.
Ana María Alderete y Edgardo Pérez tuvieron la amabilidad de leer parte del material en sus primeras etapas, agradecemos la oportunidad de sus apreciaciones y sugerencias tempranas.
Con Horacio Faas, y luego de compartir el tratamiento de los temas de probabilidad e inferencia, surgieron valiosas reflexiones sobre las modalidades del razonamiento inductivo y los límites de la formalización.
Algunos compañeros de la cátedra de Psicoestadística están mencionados por sus colaboraciones en los capítulos indicados. Pero hubo quienes hicieron aportes de importancia en las actividades prácticas de aplicación y en la lectura y corrección de originales. En esta tarea corresponde destacar al Profesor Osvaldo Bertone y a los ayudantes Marcelo Vaiman y Virginia Fornero.
Agradecemos a Carmen Díaz Batanero cuyos comentarios e interrogantes ayudaron a mejorar la forma de presentar algunos conceptos y a cuestionar la presentación de algunos temas, en especial lo relacionado a probabilidad condicional.
Al equipo InfoStat® de la Facultad de Ciencias Agropecuarias, por el desarrollo del software y su puesta a disposición de los alumnos.
Marcelo Casarín hizo la valiosa contribución de trabajar sobre los textos originales, mejorando su legibilidad.
Finalmente una mención a Eduardo Cosacov, con quien nos hemos encontrado compartiendo el interés porque nuestra facultad genere bibliografía propia, adecuada a sus alumnos y que valorice las producciones locales.

Independientemente de los aportes mencionados, los errores que se encuentren en los textos son, por cierto, de exclusiva responsabilidad del autor.

Introducción

Este manual busca ser una referencia introductoria para los procedimientos estadísticos básicos que más se usan en Psicología, Educación y otras Ciencias Sociales.

El tratamiento de los temas no supone que se haya estudiado Estadística con anterioridad, por lo que puede considerarse accesible a cualquier estudiante universitario. El texto no busca desarrollar habilidades de cálculo, sino comprensión del razonamiento que sostiene a cada procedimiento. No se trata de aprender a hacer operaciones aritméticas, porque eso se hace con un programa diseñado a ese efecto, que automatiza las cuentas. Nuestro trabajo será el de entender para qué sirve un procedimiento, cuándo corresponde usarlo y cómo se lee el resultado, ya que ese es el uso que se hará de la estadística en el desempeño profesional o en investigación.

Para quienes se inicien en los temas de estadística, al final de cada capítulo hemos provisto de actividades prácticas de repaso. Estas actividades sirven como control de lectura sobre los contenidos mínimos que deberían manejarse al cabo de cada capítulo.

Las aplicaciones informáticas se han realizado usando InfoStat®, (Di Rienzo, 2008) un paquete de análisis de datos desarrollado en la UNC y que ha sido adoptado por la Facultad de Psicología. Además de la sencillez de su operación, el programa puede obtenerse en versión estudiantil sin costo en www.infostat.com.ar. El manual de este programa (Balzarini, 2008) es un adecuado complemento para operaciones de mayor grado de complejidad que las presentadas aquí.

Por tratarse de una disciplina que tiene sus raíces en el terreno de la Matemática, no siempre la estadística resulta amena para el estudiante de Ciencias Sociales. Es por ello que conviene dedicar unas líneas a quienes puedan encarar esta materia con una disposición temerosa o despreciativa hacia lo relacionado con números.

Existen muchas razones para valorar el estudio de la estadística como materia especifica de Psicología, en Educación y en otras Ciencias Sociales. En el curso de nivelación de Psicología, por ejemplo, vemos una referencia, muy pequeña, en el esquema de "El gran árbol de la Psicología". Allí, la estadística es una regadera con la leyenda *antiparasitario*, no parece un rol central, sin dudas. Sin embargo, el autor de esta representación metafórica habrá querido indicar que la Estadística permite reducir el efecto de ideas parásitas, ideas que no se desprenden de lo que se observa, ni se derivan de ello, sino que

constituyen prejuicios, creencias previas, que contaminan lo que se observa. Toda ciencia debe superar estas creencias, esta ilusión de conocer, pero especialmente importante es eliminar los prejuicios en el campo de las Ciencias Sociales, porque es allí donde más abundan. Expresiones como "Esta persona es así porque de chico no lo tenían en cuenta", "No está dotado para la matemática", "Los sueños anuncian lo que va a pasar"; "Las mujeres tienen más sensibilidad que los hombres", son en general, falsas, provienen de creencias, de tradiciones, de voces populares transmitidas de una generación a la siguiente. La Estadística aporta a la investigación, al descubrimiento de relaciones entre hechos y a fundamentar esos descubrimientos. Lo hace con una mirada aguda y acotada, pero necesaria para "limpiar" las observaciones de la ilusión de conocer, en particular en un terreno en el que lo que observamos nos atañe de manera muy próxima, y podemos confundir "lo que sucede" con lo que creemos, opinamos o suponemos sobre ello.

Indiquemos algunas razones más inmediatas que justifican la presencia de Estadística en las carreras de Psicología y Educación.

En primer lugar, hay varias materias en la carrera que requieren que se conozca estadística básica: Metodología de la Investigación, Técnicas Psicométricas, Psicología Sanitaria; y hay muchos campos en los que la Estadística puede jugar un papel de importancia, como el de la Criminología, la Psicología Laboral o la Psicología Política. Este manual contempla las necesidades de contenidos de otras materias por lo que los conceptos que aquí se presentan volverán a verse aplicados a distintos contenidos. Por esta razón el énfasis está puesto en el aprendizaje de procedimientos y en la comprensión de razonamientos y no en la memorización de fórmulas.

En segundo lugar, quienes se dediquen al ejercicio profesional aplicarán técnicas de intervención en sus distintos campos de especialización, y estas técnicas están basadas en la teoría y en el estado del conocimiento en un momento dado. Así como luego de un tiempo de haber usado una droga puede descubrirse que no produce los efectos deseados, también vale esto para cualquier intervención profesional: terapéutica, educativa, social. Hasta hace algunas décadas se creía que golpear a los niños mejoraba su aprendizaje, cuando eso se consideró cruel, se buscaron formas de castigo menos dolorosas. Pero la investigación demostró que castigar una conducta inadecuada es, a menudo menos eficaz que premiar una adecuada, lo que condujo a grandes cambios en las recomendaciones sobre cómo educar a los niños y cómo facilitar los aprendizajes, de cualquier naturaleza.

Cuando repetimos que el conocimiento científico es revisable, queremos decir justamente eso: que en cualquier momento puede hallarse nueva evidencia que contradiga las convicciones que teníamos antes. Por cierto, no se trata de cualquier evidencia: si observamos, en un caso que una persona se enfermó luego de tomar un medicamento, no por eso estamos autorizados a solicitar que se saque de la venta al público ese medicamento. Por el contrario, la evidencia a la que nos referimos es la que se obtiene con procedimientos cuidadosos de observación, registro, comparación, medición y análisis; en pocas palabras, con investigación. Lo que sabemos y lo que aprendamos para desempeñarnos como profesionales es el conocimiento de que se dispone en este momento, y que está en continua reelaboración a través de la investigación. Una vez que quienes hoy están estudiando en la universidad terminen sus carreras y trabajen como profesionales, asistirán a cambios en el modo de intervenir, los psicólogos encontrarán nuevas terapias, los educadores nuevas estrategias didácticas. Eso no se aprenderá en la Facultad, se aprenderá luego, manteniéndose actualizado, leyendo revistas científicas, asistiendo a congresos; en fin enterándose de cómo cambia el conocimiento y se revisan los saberes a partir de los resultados de la investigación. Y la investigación usa la estadística muy a menudo. Si no se puede leer un artículo científico porque no se entiende lo que dicen las cifras, solo se podrá tener una idea general del resultado, más grave aún; puede que decidamos que solo vamos a leer la información que no contenga cifras, con lo cual accederemos a una pequeña parte de lo que sucede. O bien quedaremos aferrados a procedimientos que aprendimos una vez y que nunca revisamos.

Nuestra limitación puede ser muy peligrosa: si no entendemos cómo se obtuvo un resultado, no podremos cuestionarlos, no podremos dudar de ellos, no podremos discutir procedimientos que nos son ajenos.

Por último, y aunque no parezca obvio, la Estadística nos sirve en la vida diaria. El ejercicio de nuestros derechos ciudadanos necesita que podamos darnos cuenta de lo que nos dicen las mediciones de audiencia, las consultoras políticas, los laboratorios de medicamentos, los indicadores nutricionales de lo que comemos, las estadísticas oficiales (tasa de desempleo, pobreza), entre otras fuentes de información. ¿A quiénes consultaron para decidir que un programa de televisión se levanta y otro se sostiene? ¿Cómo se hacen las encuestas que indican quién va a ganar las elecciones? ¿Qué implica que una técnica anticonceptiva sea eficaz en el 99% de los casos? Somos nosotros los que vemos la programación que se ofrece, somos nosotros los afectados por los resultados de una elección de

autoridades, somos nosotros los que consumimos. Mucha de esa información usa estadísticas y hay parte del vocabulario que ignoramos pero que, por habernos habituado a escuchar, creemos conocer, ya que las palabras nos suenan familiares: el promedio, un porcentaje, que una diferencia sea significativa. Cada una de esas expresiones tiene un significado preciso: si no lo conocemos somos fácil presa para el engaño.

Quizás que no lo conozcamos —y que muchos prefieran no conocerlo— no sea por azar, de hecho, es una manera de disponer de consumidores que no cuestionan, que no molestan con preguntas.

Así, puede entenderse que la materia no se limite a enseñar un conjunto de técnicas, sino que más bien busque ofrecer herramientas que ayuden a mirar lo que nos rodea desde una posición más informada y de allí también más crítica.

Eduardo Bologna
Febrero de 2011

Presentación: ¿Estadística en Psicología y Educación?

Eduardo Bologna
María Marta Morales

La Estadística nos ayudará a entender comportamientos, procesos y fenómenos individuales y sociales, y lo hará desde una perspectiva que puede parecer un tanto ajena a quienes se inician en el estudio de carreras de Ciencias Sociales. Cuando se aplica a fenómenos sociales, la Estadística cumple la función de tomar distancia de aquello que se observa. La Estadística no analiza individuos aisladamente, sino conjuntos de ellos, conjuntos a los que define de acuerdo a ciertas características que elige deliberadamente. Grupos de personas de determinada edad, clase social, nivel de educación, nivel de inteligencia, hábitos, etc. y es entre esos grupos que hace comparaciones y busca similitudes y diferencias. Usa estas clasificaciones y las comparaciones entre los grupos a fin de identificar factores que expliquen las diferencias entre individuos.

¿Cómo puede aportar la Estadística a la Psicología, si ésta es ciencia de lo particular, si cada persona es única? ¿Qué de la especificidad de cada experiencia individual de aprendizaje? Para entender esto se debe, en primer lugar, recordar que cada hecho social o individual que se considera y que demanda explicación, está determinado por un conjunto de factores muy amplio, se trata de lo que llamamos multicausalidad. Es decir, no hay hechos psicológicos, educativos, ni sociales que puedan explicarse a partir de una única "causa". En segundo lugar, el conjunto de factores que explican un hecho, puede dividirse en aquellos que afectan al individuo de modo exclusivo y aquellos de carácter colectivo. Los individuales hacen de cada sujeto un caso único, los colectivos ubican al individuo en similitud con quienes comparte un grupo o diversos grupos. Veamos esto en ejemplos: La ansiedad de una persona particular frente a una situación, se explica por muchos factores, algunos de ellos son individuales y otros son generales. Una entrevista laboral o un examen oral son situaciones que generan más ansiedad que una conversación entre amigos y esto es así para casi cualquier persona. Pero para algunos esa ansiedad es leve y soportable y a otros les

dificulta un desempeño de buena calidad. El carácter ansiógeno de la situación puede analizarse de manera colectiva, preguntándonos si los varones tienden a mostrar más ansiedad que las mujeres ante esa situación o si las personas más exigentes consigo mismas sufren de mayor ansiedad. Es de este modo que explicamos *parte* de las diferencias en el nivel de ansiedad, por factores generales, otra *parte* será explicada por características del individuo, que lo hacen único.
La dificultad de un niño para estudiar Matemática puede explicarse por el modo en que la materia se enseña, por la actitud de sus padres hacia la matemática, por sus propias creencias acerca de la dificultad intrínseca de la materia y por otros factores más íntimos relacionados con su historia personal. Algunas de estas explicaciones pueden analizarse de manera colectiva, buscando mejores métodos para enseñar Matemática, investigando de qué manera la actitud de los padres o las creencias de los estudiantes pueden incidir en su desempeño en la materia. El estudio de estos factores colectivos se ve auxiliado por la Estadística, que permite tratar con conjuntos de individuos y ver las regularidades que solo se aprecian cuando se los considera agrupados.

Veamos cómo se materializa este cambio en la mirada desde el individuo hacia el grupo. La siguiente es una lista de las materias que tienen aprobadas algunos alumnos de segundo año de una carrera universitaria:

Alumno	Materias aprobadas
Susana	5
Marcos	6
Daniel	5
Federico	4
María	4
Pedro	5
Eugenia	5
Mabel	5
Francisco	5

La lista los individualiza, los reconoce por su nombre, nos dice cuántas materias aprobó cada uno. Si transformamos esa lista en una tabla:

Materias aprobadas	Cantidad de alumnos
4	2
5	6
6	1

Leemos ahora que con cuatro materias aprobadas hay dos alumnos, con cinco hay seis y solo uno tiene seis materias aprobadas. Las personas desaparecieron, ya no hay nombres, hemos abstraído para referirnos a las *materias aprobadas*, no a los alumnos. En la tabla vemos que lo más frecuente es que tengan cinco materias aprobadas y que seis es excepcional. Hemos pasado de la lista de individuos a la tabla de valores. Nos despegamos de los casos a fin de buscar la regularidad en el conjunto.
Eso hace la Estadística, es una operación muy importante en la Psicología y en Educación, porque está dirigida a ver el modo en que los factores generales afectan lo que se observa, más allá de los casos particulares, para después volver al caso individual.
Esas generalizaciones son las que permiten, por ejemplo, recomendar un tipo de intervención terapéutica y desalentar otras. O bien, si sabemos que lo normal es esperar que un niño comience a hablar entre los 12 y los 24 meses, es porque muchos niños han sido observados y ha podido establecerse esa regularidad. Conocer eso nos permitirá saber que si un niño en particular, a los tres años de edad no habla, necesita alguna intervención específica. O, en otro ejemplo, si hemos podido verificar que las personas que se proponen lograr metas muy elevadas tienden a sentirse más ansiosas cuando son evaluadas, podremos intervenir, ante un caso particular, sobre la fijación de metas a fin de reducir la ansiedad.
Conocer las regularidades grupales no implica dejar de lado al individuo, por el contrario, implica situarlo en relación a un grupo y conocer factores de orden general que pueden estar afectándolo individualmente. Esto se combina, en cada caso, con la historia subjetiva para dar lugar al carácter único de cada persona.
Los datos agregados sobre patologías psicosociales como el suicidio o la depresión pueden mostrarnos las tendencias y ayudarnos a decidir a qué es necesario atender con mayor urgencia. El análisis de esas tendencias por edades, o por clase social, permite identificar grupos especialmente vulnerables.

En Educación, el campo de aplicación de la Estadística es tanto o más vasto que en Psicología, porque los organismos gubernamentales y no gubernamentales producen grandes cantidades de datos, muy valiosos para detectar problemas de determinados sectores de la población, o para comparar entre países, entre sistemas educativos diferentes.

Hay, en quienes trabajan desde la estadística aplicada en las Ciencias Sociales, una especie de desapego, de alejamiento de lo particular. En

Pensamientos, VII, Platón señala: "Para hablar de los hombres, es necesario examinar las cosas terrestres como desde un lugar elevado, las organizaciones y las expediciones armadas, las uniones y las rupturas, los nacimientos y las muertes, el tumulto de las tribunas y los campos desiertos, la diversidad de naciones, las fiestas y los duelos, los mercados, las mezclas y los contrastes; y ver el orden que de allí nace". Pierre Hadot (2007) muestra que en las escuelas filosóficas, los neófitos debían aprender a suspender todo juicio de valor y toda proyección afectiva, ejercitarse en tomar altura, en el sentido más material de la expresión, imaginando que volaban sobre la tierra, observándola como un objeto lejano. Alcanzado ese punto de vista, podían extraer, del desorden aparente de las cuestiones humanas, la regularidad de un orden universal y divino. Herrán (2002) señala que esta cultura del desapego jugó un rol central en el desarrollo del espíritu científico.

Creencias sobre la Estadística

A veces hay quienes creen que la introducción de procedimientos estadísticos en la investigación aporta objetividad al análisis. O bien que el investigador que usa técnicas estadísticas busca cierta asepsia en el conocimiento que construye, busca dotarlo de veracidad más allá de las argumentaciones. Como si la introducción de números en el discurso lo volviera más serio o más válido. Se trata de creencias equivocadas del sentido común, poco informado, ya que la Estadística solo puede aportar rigor al análisis de observaciones de buena calidad y solo nos ayuda en la organización y en la posibilidad de generalizar nuestros resultados. Sirve para poner en correspondencia las ideas con lo que se observa, la teoría con los hechos.

Sin embargo, en un uso malintencionado, la Estadística puede ponerse al servicio de "probar" falacias que, a la vista de quienes no pueden interpretar los resultados, aparecen como verdades irrefutables[1]. Aprender Estadística debe servirnos también para desterrar el mito que consiste en creer que usar técnicas cuantitativas puede dotar de cientificidad a un argumento vacío.

También puede confundirse el uso de estadísticas con una asimilación de la investigación social a la de las ciencias naturales. Si bien la Estadística permite tomar distancia de los casos individuales para poner el acento en características compartidas, en las ciencias

[1] ¿Hemos notado que los anuncios de productos cuya eficacia aparece "científicamente demostrada", siempre usan porcentajes con decimales? Parece que eso da más realismo a las cifras.

sociales eso es solo una parte de la explicación de los fenómenos, aquella parte que es compartida por los integrantes del grupo, siempre queda una componente de variabilidad que solo puede explicarse de manera individual. Aunque se puedan usar procedimientos estadísticos similares para analizar el diferente rendimiento de semillas y el rendimiento académico de los estudiantes, eso no implica que desconozcamos la complejidad de los fenómenos sociales ni el carácter único de los sujetos de nuestros estudios.

Las áreas de la Estadística

Dos amplios conjuntos de procedimientos constituyen el aporte que la Estadística hace a la construcción de conocimiento: el resumen de un conjunto grande de información y la extensión de las conclusiones que se observan en ciertos sujetos, a otros sujetos que no han sido observados. Estas dos grandes funciones son consideradas como dos áreas de la Estadística, a las que se denomina descriptiva e inferencial respectivamente.

Antes de señalar los objetivos que se persiguen en cada una de estas áreas, conviene hacer una breve referencia a la distinción entre muestra y población —aunque volveremos sobre ella en el capítulo dedicado al *muestreo*—, ya que ella articula las dos áreas. Los datos que recogemos están limitados en su alcance a las posibilidades de nuestro estudio; si el objetivo es generalizar los resultados, esos casos que observemos serán una fracción de un universo mayor. Cuando analizamos la ansiedad frente a los exámenes en estudiantes universitarios, no está dentro de nuestras posibilidades observar a todos los estudiantes universitarios. Por el contrario, habremos de seleccionar a algunos de ellos. Ese conjunto de estudiantes que seleccionamos se llama *muestra* y nos aportará la información sobre la que trabajaremos. Según el modo en que la muestra haya sido seleccionada, los resultados se podrán extender a una *población de referencia*, que en este ejemplo podría ser la del total de estudiantes de una universidad (de la que se extrajo la muestra). Las condiciones para que esta generalización sea posible serán expuestas en detalle en el capítulo correspondiente.

La descripción estadística

Debido a que en estadística no trabajamos con individuos aislados sino con conjuntos de ellos, siempre es necesario resumir la información, para presentarla de manera accesible a la lectura y para extraer significado.

Una gran tabla proveniente de registros hospitalarios que muestre las edades de madres primerizas no puede leerse de manera directa; es necesario buscar indicadores de síntesis, uno de ellos, muy frecuente, es el promedio. Sucede del mismo modo si contamos con los puntajes de una prueba de memoria aplicada a muchas personas. O también si tenemos el listado de alumnos de una escuela y de cada uno sabemos si repite o no el curso. En esos casos podemos resumir esa información indicando el promedio (con las limitaciones que esta medida tiene, como veremos en el capítulo 3). También es posible indicar cuántas personas tienen un valor menor a cierta cifra o mayor a otra: ¿Cuántas de las madres primerizas son menores de 20 años?, ¿repiten de grado con igual frecuencia los varones que las mujeres?. O bien expresar los valores a través de gráficos, que suelen aportar mucha información de manera abreviada (aunque a veces también pueden ser engañosos).
Cuando de cada alumno de una escuela conocemos si repite el curso o no, es conveniente calcular una tasa de repitencia para cada curso o para cada escuela, en lugar de indicar la condición de repitente o no de cada alumno.

Así, la Estadística descriptiva nos proveerá de una serie de procedimientos dirigidos a resumir, a sintetizar información, a volverla manejable para que podamos interpretarla y extraer conclusiones a partir del conjunto de datos que, de otra manera, serían ininteligibles.

La inferencia estadística
Una vez que disponemos de una síntesis de la información que hemos recogido de un conjunto de individuos, nos interesa otro problema: el de preguntarnos si eso que observamos vale también para otros, a los que no hemos observado. Si hemos visto a muchas personas, algunas exigentes consigo mismo y otras que no lo son y hallamos que las primeras manifiestan más ansiedad en los exámenes que las segundas, ¿podemos decir que la autoexigencia incide en la ansiedad?, es decir, ¿podemos generalizar nuestro resultado? La Estadística inferencial se ocupará de esto, de decirnos bajo qué condiciones se pueden extender nuestros hallazgos a casos no observados.
Si encontramos niños de madres que han tomado bebidas alcohólicas durante el embarazo y niños de madres que no bebieron y descubrimos que, en promedio, los hijos de madres bebedoras tienen niveles de desarrollo motor más bajo que los hijos de madres no bebedoras ¿podemos afirmar que beber alcohol durante el embarazo retrasa el desarrollo motor de los hijos? Según cuántos niños

hayamos observado, según qué tan grande sea la diferencia entre el promedio de desarrollo motor de los hijos de bebedoras y no bebedoras, según qué tan variable sea el desarrollo entre los niños, tendremos o no argumentos para generalizar el resultado y afirmar que existe o no una relación entre consumo de alcohol durante el embarazo y desarrollo motor del niño.
En todos los casos, cuando sea posible hacer generalizaciones, éstas estarán limitadas a un contexto específico. El análisis que se haga, de la relación entre pobreza y educación en Argentina, puede no ser válido para la población de Brasil. Es decir, debe estar explícita cuál es la población de referencia a la cual es válido extender los resultados que se obtienen.

Puede considerarse a la descripción como una etapa anterior a la inferencia, ya que esta última no puede lograrse sin una adecuada descripción previa de la información. Pero la descripción tiene entidad propia y, como se verá en Metodología de la Investigación, un estudio puede plantearse objetivos exclusivamente descriptivos, que no se dirijan a generalizar los resultados obtenidos.

La observación de regularidades a escala colectiva puede ser también un medio para plantear preguntas de investigación. Fue éste el camino seguido por Durkheim (1994 [1897]) quien, al observar que las tasas de suicidio de diferentes comunidades aparecían muy disímiles, postuló que, más allá de las razones particularísimas que cada persona podría tener para llegar al suicidio, debía haber otros factores, de orden social, que determinaran los suicidios. Va así a preguntarse qué elementos distintivos de las comunidades pueden explicar que en unas el suicidio sea más frecuente que en otras.
Si apreciamos, por ejemplo, que el rendimiento de los alumnos de escuelas urbano periféricas es menor que el de los alumnos de escuelas urbanas, nos preguntaremos ¿qué hace que se produzca esa diferencia? O en el ejemplo anterior, sobre las madres que beben durante el embarazo ¿cuáles son las etapas del embarazo en que la ingesta de alcohol es más peligrosa? o ¿cuál es el mecanismo fisiológico que liga el alcohol ingerido por la madre con el desarrollo motor del niño?
En estos ejemplos, el resultado estadístico se halla en el origen de la indagación, en la construcción del problema de investigación y no (solamente) en el análisis posterior de los datos recogidos. Esto muestra el carácter herramental de la Estadística: su uso siempre estará al servicio de la producción de conocimiento validado.

Capítulo 1: Las variables y su nivel de medición

María Cristina González de Menne
Ana Eugenia Faas
Eduardo Bologna

"Esas ambigüedades, redundancias y deficiencias recuerdan las que el doctor Franz Kuhn atribuye a cierta enciclopedia china que se titula *Emporio celestial de conocimientos benévolos*. En sus remotas páginas está escrito que los animales se dividen en (a) pertenecientes al Emperador, (b) embalsamados, (c) amaestrados, (d) lechones, (e) sirenas, (f) fabulosos, (g) perros sueltos, (h) incluidos en esta clasificación, (i) que se agitan como locos, (j) innumerables, (k) dibujados con un pincel finísimo de pelo de camello, (1) etcétera, (m) que acaban de romper el jarrón, (n) que de lejos parecen moscas."
Jorge Luis Borges, *El idioma analítico de John Wilkins*

La ciencia se interesa por la producción de conocimiento validado, uno de cuyos requisitos es la objetividad. Sin perder de vista que tanto los criterios de validación como el concepto mismo de objetividad son motivo de debate epistemológico, partiremos en este curso de la necesidad de usar un lenguaje que pueda intercambiarse entre investigadores y que dependa, en el menor grado posible, de las impresiones subjetivas de cada investigador individual.

Un modo de acercarse a lograr esta comunicabilidad de las ideas y de los resultados de observaciones es definiendo de la manera más precisa que sea posible los elementos acerca de los que se habla.

Es frecuente la expresión "esta persona es más inteligente que aquella". ¿Qué queremos decir exactamente con eso?, la afirmación podría provenir de algún evento en que vimos a esa persona actuando de manera que llamaríamos inteligente, aunque esto también puede confundirse con astucia: no es infrecuente usar el adjetivo inteligente para un estafador, alguien a quien le resulta fácil engañar a otros; y, a la inversa, sería poco inteligente quien se deja engañar con facilidad. O bien, a menudo decimos que alguien es inteligente porque obtiene buenos resultados en sus estudios. Se observa que contar con una definición de inteligencia permitirá decidir cuándo aplicar esa idea a alguien, cuándo una conducta es

inteligente, cómo desarrollar la inteligencia. Si se puede definir el concepto con el que se trabaja, se pueden indicar ciertas operaciones a realizar para evaluarlo en cada caso particular. Así, si definimos la inteligencia como la capacidad para resolver problemas, podremos diseñar un conjunto de problemas, cada uno más difícil que el anterior, y observar cuántos de esos problemas puede resolver una persona; si otra persona resuelve un número mayor de ellos, estaremos autorizados para decir que es más inteligente. No es diferente esta operación al trabajo que realiza el oftalmólogo cuando solicita que se lean, desde cierta distancia, líneas de letras y símbolos de tamaño cada vez menor. Él evaluará la visión de acuerdo a cuántas líneas de letras y símbolos llegue a ver el paciente: tanto mayor será su capacidad visual cuantas más líneas alcance a ver.

Así, vemos que antes que nada es necesario definir el concepto con el que se trabaja, luego se requiere diseñar un instrumento que refleje esa definición y finalmente aplicar este instrumento a las personas que se evaluarán. Al hacer esto último se obtiene un valor que, si se expresa de manera cuantitativa, permite hacer comparaciones de ese concepto entre personas, entre grupos, etc.

¿Podemos comparar personas? La respuesta es no, por el contrario lo que sí pueden compararse son características claramente definidas de las personas. Del mismo modo no se pueden comparar escuelas, ni hogares, ni países si no se especifica en qué aspecto se realiza la comparación. O dicho de otro modo, cuál es la característica que se compara, y cómo se mide esa característica.

Podemos decir que una persona tiene más escolarización formal que otra, indicando con eso que ha aprobado más años de la escuela o de la universidad. Podemos decir que un hogar es diferente a otro si uno se compone de una pareja sola y el otro incluye tres hijos. Un país puede tener más habitantes, un régimen político diferente, o mayor libertad de expresión que otro. En todos los casos especificamos una característica, un rasgo sobre la base del cual hacemos la comparación.

Vamos a introducir ahora dos definiciones para los elementos que hemos mencionado hasta aquí.

En primer lugar, veamos que los entes cuyos aspectos se comparan pueden ser diferentes: personas, hogares, países, escuelas, etc. Esas entidades se llaman **unidades de análisis** (a menudo indicadas UA). Son los elementos entre los que se compara alguna cualidad, son los sujetos o individuos, de manera general. Así, en la afirmación “en la escuelas-urbano

periféricas hay más alumnos que repiten curso que en las urbanas", las unidades de análisis son las escuelas. Si se afirma que "las personas de menores recursos acceden menos frecuentemente a la educación superior", hablamos de personas, y estas son las unidades de análisis. Y es muy diferente a esta otra afirmación: "en los países más pobres, es menor la proporción de personas que acceden a la educación superior", en la que las unidades de análisis son los países.

Se llama **unidades de análisis** a los entes individuales acerca de los que se analizan sus cualidades.

En segundo lugar, hay algo que se compara: la inteligencia, la composición del hogar, el régimen político, el número de habitantes. Estas son las características de las unidades de análisis que se someten a comparación, se denominan **variables**. Las variables son los aspectos de los individuos que se someten a comparación; su cualidad central es la que le da nombre: la de variar.

Llamamos **variable** a una característica de las unidades de análisis que puede asumir diferentes valores en cada una de ellas.

Cada vez que se haga referencia a una variable, debe conocerse cuál es la unidad de análisis a la que se refiere, si no resulta claro, se debe indicar. Es diferente afirmar que un país es rico que decir que sus habitantes lo son.

En cada individuo (en cada unidad de análisis) la variable asume un valor que puede ser el mismo o diferente del de otro individuo. Así, la edad puede asumir el valor "21" para una persona y "20" para otra, el régimen político puede ser "democracia presidencialista" en un país y "monarquía parlamentaria" en otro. Una persona puede tardar 2 segundos en reconocer una imagen y otra tardar 2,5 segundos; allí diremos que la variable "tiempo para reconocer la imagen, expresado en segundos" asume el valor 2 para la primera persona y 2,5 para la segunda.

Por el contrario, si una cualidad es la misma para todas las unidades de análisis, no es posible ninguna comparación. Por ejemplo, si nuestro universo está compuesto por estudiantes universitarios, no podemos comparar el "nivel de educación", ya que todos ellos tienen el mismo. Para ese universo, el nivel de

educación no es variable, diremos por lo tanto que carece de variabilidad[2].

En tercer lugar, hemos dicho que las variables asumen valores para cada unidad de análisis, esos valores a menudo se denominan categorías.

Son **categorías** de una variable los valores que puede asumir.

Cuando se define una variable debe indicarse también el conjunto de categorías que le corresponden, aunque a veces esto está implícito. Si la variable es *sexo*, las categorías son varón y mujer, si se trata del *nivel de escolaridad alcanzado*, pueden considerarse las siguientes categorías: ninguno, primario incompleto, primario completo, secundario incompleto, secundario completo, terciario o universitario incompleto, terciario o universitario completo y postgrado. Si tratamos con la variable *edad*, sus categorías son valores numéricos, entre cero y 100 años.

Hay dos propiedades que debemos asegurar que cumplan las categorías que construyamos. La primera se llama **exclusión mutua**, es decir que cada categoría excluya a todas las demás. Dicho de otra manera, si a un individuo le corresponde una categoría, entonces sabemos que no le corresponde ninguna otra. Si analizamos hogares y a cada persona le preguntamos por su parentesco, sin indicar con quién, tendremos una categorización defectuosa, porque una persona del hogar puede al mismo tiempo ser hijo y padre, si conviven tres generaciones, o hijo y hermano. De este modo a la misma persona le corresponderían dos categorías y se viola el requisito de exclusión mutua. Esto se resuelve estableciendo respecto de quién se declara el parentesco, y todos los integrantes del hogar lo refieren a la misma persona[3].

Al analizar los *tipos de lectura preferida*, nos equivocaríamos si categorizáramos como de ficción, de misterio, policiales, románticas, biográficas, de aventuras, ya que la categoría ficción puede incluir misterio, policiales o novelas románticas o de aventuras.

---

[2] Lo opuesto a variable, es constante. Una constante es un valor igual para todas las unidades que se analizan. En nuestras ciencias no son muy frecuentes, pero sí en las naturales, por ejemplo el punto de fusión del hielo (a presión atmosférica) es una constante.

[3] Habitualmente denominado "Jefe de Hogar".

También se comete ese error si se clasifica a las escuelas como céntricas, parroquiales, urbanas, rurales. Dado que una escuela puede ser al mismo tiempo parroquial y urbana. Es necesario separar, para que quede claro, lo que interesa en el análisis: si lo que queremos distinguir son escuelas céntricas de barriales, entonces la variable será la ubicación geográfica y no implica el carácter parroquial o no de la escuela.
Resulta muy importante que, una vez definida la variable, se verifique que sus categorías sean mutuamente excluyentes, porque de lo contrario, cuando se observa, no queda claro cómo hacer la clasificación y diferentes observadores clasificarán de manera distinta el mismo hecho.

Las categorías de una variable son **mutuamente excluyentes** si a cada individuo le corresponde no más de una categoría.

El segundo requisito que solicitaremos a las categorías de una variable es que agoten todas las posibilidades de variación, es decir, que todos los valores posibles estén contemplados. Esta cualidad se llama **exhaustividad**.
Veamos qué sucede si no respetamos este requisito. Si evaluamos la variable situación conyugal y ofrecemos como categorías: casado, soltero, divorciado, viudo; las personas que estén viviendo juntas sin estar casadas no encuentran un lugar donde ubicarse, como tampoco lo encuentran quienes están separados sin haberse divorciado. Para resolver esto es necesario, o bien incluir estas categorías separadamente: casado, unido, soltero, separado, divorciado, viudo; ampliando así el número de categorías, o bien fusionándolas con las existentes: casado o unido, soltero, separado o divorciado, viudo. Esta última fue la opción elegida en el ejemplo del nivel de educación, que mostramos más arriba, en cuya categorización fusionamos terciario y universitario.
Cuando mencionamos el ejemplo de la edad, vimos que las categorías son valores numéricos que pueden ir del cero hasta el 100, pero ¿qué sucede con las personas que alcanzaron una edad superior a 100 años? Quizás sean pocas, pero no pueden quedar sin categoría donde incluirse. Por lo demás puede haber una de 103 años, otra de 105, y no se justifica seguir extendiendo categorías. Una solución frecuente es la de tomar una categoría "abierta final", fijando como última categoría 100 y más, e incluir allí a todas las personas que declaren una edad de 100 años o superior. Puede verse que esta opción conlleva una pérdida de información, ya que no sabemos la edad exacta de quienes se

ubican en esa categoría. Aceptamos esa pérdida a cambio de reducir el número de categorías de la variable, y de esto trataremos en el punto siguiente. Sin embargo, cuando usemos las distribuciones de frecuencia para calcular medidas descriptivas (capítulo 3, en especial en el cálculo de la media) no será posible apelar a este procedimiento para reducir las categorías.

Seguramente hemos observado en cuestionarios que, luego de un conjunto de opciones para responder, se incluye una categoría que dice "Otro… especificar". Se trata de casos muy interesantes de categorizaciones en las que no se sabe de antemano cuáles son todas las respuestas posibles; son frecuentes en las encuestas de opinión. Por ejemplo, si alguien declara que en las próximas elecciones va a votar en blanco y preguntamos por qué, podemos conocer de antemano algunas de las respuestas posibles, pero debemos dejar espacio para que los encuestados expresen razones que no habíamos previsto. De este modo aseguramos la exhaustividad de las categorías.

Las categorías de una variable son **exhaustivas** si todo individuo tiene alguna categoría que le corresponda.

En algunas situaciones, el número de categorías de una variable es parte de nuestra decisión. Hay casos en que las categorías están establecidas de antemano: por ejemplo, en la variable sexo tendemos a usar como categorías las de varón y mujer; sin embargo, si estamos frente a un estudio que trate precisamente sobre orientación sexual de las personas, deberán considerarse otras categorías.

Cuando mencionamos el ejemplo de la edad de las personas, vimos que es nuestra elección terminar las categorías con 100 y más. De hecho, también podríamos haber mantenido las edades exactas hasta 109 años y cerrar con 110 y más. Es nuestra elección y depende de cuánta información y cuánta claridad decidamos que tenga nuestra clasificación; lamentablemente, no es posible lograr al mismo tiempo el máximo de información y de claridad en la presentación[4].

Veamos dos opciones para el caso de nivel de educación, según se elija fusionar o no los estudios terciarios con los universitarios:

[4] Veremos muy a menudo que, en Estadística, es necesario llegar a puntos de equilibrio entre el grado de detalle de la información que se ofrece y la claridad con que esa información puede presentarse.

Tabla 1: Ejemplo de diferentes categorizaciones para la variable *Máximo nivel de educación formal alcanzado*

Máximo nivel de educación formal alcanzado
ninguno
primario incompleto
primario completo
secundario incompleto
secundario completo
terciario o universitario incompleto
terciario o universitario completo
postgrado

Máximo nivel de educación formal alcanzado
ninguno
primario incompleto
primario completo
secundario incompleto
secundario completo
terciario incompleto
terciario completo
universitario incompleto
universitario completo
postgrado

La decisión sobre cuál de las dos categorizaciones es preferible está a cargo del investigador. Así, si no es de interés distinguir terciario de universitario, el primer conjunto de valores es más conveniente, porque tiene menos categorías y es más simple para presentar.
En el próximo capítulo nos ocuparemos de la opción de reducir la cantidad de categorías por medio de la agrupación de valores numéricos. Por ejemplo, en lugar de tomar la edad exacta de las personas, es posible establecer como categorías a conjuntos de valores: de 0 a 9, de 10 a 19, etc.

El uso de símbolos numéricos

Como hemos visto, las categorías pueden tener diferente naturaleza: algunas se expresan con números (como la edad) y otras con palabras (como el tipo de hogar); sin embargo es muy común representar con números a las categorías, aun cuando lo que se observe no sea numérico. Así, en la primera categorización de la variable de la tabla 1, podemos codificar las categorías de la siguiente manera:

Tabla 2: Codificación numérica de las categorías de la variable *Máximo nivel de educación formal alcanzado*

Código	Máximo nivel de educación formal alcanzado
1	ninguno
2	primario incompleto
3	primario completo
4	secundario incompleto
5	secundario completo
6	terciario o universitario incompleto
7	terciario o universitario completo
8	postgrado

Hemos usado números para referirnos a las categorías a fin de simplificar la notación. Cuando usemos un programa informático para analizar los datos, veremos que esta codificación es necesaria.
De manera equivalente podemos codificar las categorías de otras variables:

Tabla 3: Ejemplo de codificación numérica de la variable *Sexo*

Código	sexo
1	varón
2	mujer

Tabla 4: Ejemplo de codificación numérica de la variable *Opinión sobre una propuesta de cambio de horarios de clase*

Código	Opinión sobre una propuesta de cambio de horarios de clase
1	Completamente de acuerdo
2	De acuerdo
3	Indiferente
4	En desacuerdo
5	Completamente en desacuerdo

En las variables cuyas categorías son numéricas, no es necesario hacer ninguna codificación. Así, la edad quedará expresada de manera numérica directamente por la cantidad de años. El tiempo en reconocer una imagen se medirá en el número de segundos, sin codificación[5].

[5] En el próximo capítulo veremos que es posible codificar una variable cuyas categorías son números, si lo que se desea es agrupar valores: por ejemplo, en lugar de tomar el valor numérico del ingreso (expresado en pesos) es posible crear categorías "ingresos bajos", "ingresos medios" e "ingresos altos" y codificarlas, por ejemplo como 1, 2 y 3. Sin embargo, esto significa una importante pérdida de información y de

Variables y medición

En nuestra disciplina, y con mayor intensidad en la Psicometría es de plena actualidad el debate acerca de las posibilidades de medición de los fenómenos que estudiamos. Buena parte del debate gira en torno a una definición de medición, ya que según qué sea lo que se considere como tal, lo que hacemos será medir o no. La posición más tradicional corresponde a lo que el sentido común trata como medición: la estatura, las distancias, el peso, etc. Esta definición demanda algunas propiedades a las mediciones para considerarlas como tales. Se conoce como teoría clásica de la medición, y desde ese punto de vista sería muy difícil realizar mediciones en Psicología. Una definición menos restrictiva es la que propuso Stevens (1951), según la cual "medir es asignar números a los objetos según cierta regla, de manera que los números asignados en la medición, no representan propiamente cantidades, sino relaciones".

Esta última definición, basada en la teoría representacional de la medición, es la que adoptaremos en este curso aunque, como señalamos antes, la discusión sigue vigente. Desde esta definición, evaluar una variable para una unidad de análisis dada, equivale a medir esa unidad de análisis en el aspecto que la variable expresa.

Aun cuando adoptemos una definición amplia de lo que es medir, podemos intuir que no se mide una opinión del mismo modo que se mide el salario o la estatura. Esto sugiere que, dentro de las variables de las que hemos hablado hasta aquí habrá que reconocer diferencias, y estas diferencias vendrán dadas por el significado que tengan los números que asignamos a las categorías, es decir, por las reglas que ligan los números con lo que se observa.

El **nivel de medición** de una variable está determinado por el significado que tengan los símbolos numéricos que se asignan a las categorías.

Antes de avanzar en una clasificación de las variables según su nivel de medición, detengámonos a pensar en que hay una graduación en el significado que tienen los números. En la variable sexo, haber elegido 1 para varones y 2 para mujeres es de una arbitrariedad total (que podría inclusive dar lugar a quejas). Si la codificación hubiese sido al revés, habría estado

posibilidades de análisis de los datos, por lo que solo debe recurrirse a este procedimiento cuando esté claramente justificado.

igual de bien, y también lo habría estado usar el número 25 para representar a los varones y el 38 para las mujeres, aunque esto resulta un poco incómodo. Por el contrario, en la variable edad, asignar 20 a quien tiene 20 años, parece totalmente natural ¿qué otro número podríamos haber asignado? ¿Qué sucede con el nivel de educación? En el ejemplo elegimos numerar las categorías del 1 al 8; habría habido otras opciones, por ejemplo usar solo números pares o números impares u otra secuencia arbitraria, pero hay algo importante que cualquier secuencia que elijamos deberá respetar: las categorías de la variable siguen un orden y los números deben reflejarlo; no habría sido correcto usar números que no vayan aumentando, como lo hacen los niveles de educación.
Así entonces, podríamos decir que hay grados diferentes en la libertad que tenemos para asignar los números a las categorías. Esas diferencias serán el tema del apartado siguiente.

Niveles de medición
Según la mayor o menor arbitrariedad que exista en la relación que liga los números a las categorías, hablaremos de niveles de medición. Una forma diferente de decirlo es que, según cuánta restricción haya en la asignación de los números a las categorías, será el nivel de medición de las variables. Si los números se asignan de manera totalmente arbitraria, el nivel de medición se llamará **nominal** (como en la variable sexo); si los números deben respetar el orden de las categorías (como en la educación), el nivel de la variable se llama **ordinal**. Por ahora, nos detenemos en estos dos niveles.

El nivel nominal
Es el nivel más elemental de medición: las variables de este nivel tienen categorías que son solo nombres (de allí que se llamen nominales). La asignación de códigos numéricos cumple la función de designar las categorías, es decir, de distinguirlas una de otras. Además del ejemplo de sexo, podemos mencionar: tipo de hogar (Unidades de Análisis = hogares), facultad en que está inscripto (UA = estudiantes universitarios), área de especialización preferida (UA = estudiantes de Psicología); cuyas codificaciones podrían ser:

Tabla 5: Ejemplos de codificación de variables de nivel nominal

Código	Tipo de hogar
1	Unipersonal[6]
2	Monoparental[7]
3	Nuclear[8]
4	Extendido[9]
5	Compuesto[10]

Código	Área
1	Clínica
2	Educacional
3	Jurídica
4	Laboral
5	Sanitaria
6	Social

Código	Facultad
1	Psicología
2	Filosofía
3	Medicina
4	Otras

Solo por comodidad, hemos elegido codificar desde el 1 y correlativamente, no hay ninguna limitación para, por ejemplo, haber codificado el tipo de hogar del siguiente modo:

Tabla 6: Ejemplo de codificación "excéntrica" de una variable nominal

Código	Tipo de hogar
10	Unipersonal
32	Monoparental
4	Nuclear
45	Extendido
322	Compuesto

Decimos codificación excéntrica, porque es completamente inusual hacerlo de este modo, ya que solo introduce complicaciones: nadie codificaría de esta manera, aunque no es incorrecto. Pero nos interesa llamar la atención en el carácter totalmente arbitrario de la asignación de los números.

Aun con esta amplia libertad para elegir los códigos numéricos, hay algo que no podemos hacer: no es válido usar el mismo número más de una vez. Si hiciéramos esto, confundiríamos las categorías que corresponden a cada individuo. Así, si un hogar es de tipo

[6] Solo una persona
[7] Madre o padre con hijo(s)
[8] Madre y padre con hijo (s)
[9] Con otros parientes conviviendo
[10] Con otros no parientes conviviendo

unipersonal, le corresponde (según la tabla 5) el código 1, no podría usarse ese mismo número también para los hogares de tipo monoparental. Diremos que la condición que deben cumplir los números en este nivel de medición es que a categorías diferentes correspondan números distintos.
Entonces, lo que debemos recordar de este nivel de medición es que a cada categoría podemos asignarle, de manera arbitraria, uno y solo un número. Dado que esta forma de asignar los valores numéricos solo implica que éstos designan las categorías (las distinguen a una de otra), no es posible tratarlos como números en cuanto a sus propiedades aritméticas. En particular no puede sumárselos: nada puede significar que se sumen, en el tipo de hogar de la tabla 5, los códigos 1 y 2.

Una variable está medida a nivel **nominal** si los números que representan cada categoría son asignados de manera arbitraria y solo cumplen con la función de designar y distinguir categorías diferentes

Para unidades de análisis medidas a través de una variable de nivel nominal, es posible saber si corresponden a la misma categoría o a una diferente, es decir si tienen la misma cualidad (o atributo) o una diferente.
Consideremos el siguiente ejemplo, sea la variable tipo de escuela, clasificada según su ubicación:

Tabla 7: Ejemplo de codificación de la variable nominal Tipo de escuela

Código	Tipo de escuela
1	urbana
2	urbano periférica
3	rural

Si a una escuela le corresponde el número 1 y a otra también, solo podemos decir que ambas son del mismo tipo (urbanas), si a una le corresponde el 1 y a otra el 3, sabremos que la primera es urbana y la segunda rural. El hecho que el número 3 sea más grande que 2, no tiene ninguna interpretación en este nivel de medición, como, por cierto tampoco la tiene que 3 sea el triple de 1.

Si 1 y 2 son dos categorías de una variable medida a nivel nominal, el único tipo de relación que puede establecerse entre ellas es $1 \neq 2$.

El nivel ordinal

Aquí subimos un nivel, ya que a los números que solo tienen la función de designar en las variables nominales, se agrega otra función: la de reflejar el orden entre las categorías.

Simplemente ahora se trata de variables cuyas categorías indican alguna cualidad de las unidades de análisis que crece en una dirección. Eso equivale a decir que se pueden hacer entre ellas, juicios de orden, tales como una categoría es mayor que otra, una categoría es menor que otra. El ejemplo de los niveles de educación cumple con ese requisito: efectivamente, el "primario incompleto" es un nivel de estudios superior a "ninguno", pero inferior a "primario completo".

Los valores numéricos que representan las categorías rescatan ahora una propiedad adicional: el orden. Además de poder distinguir si dos sujetos tienen la misma característica analizada o una distinta como en el nivel nominal, ahora también podemos saber si un individuo (una unidad de análisis) tiene esa característica en mayor o menor grado. Así como "ninguno" es menor que "primario incompleto", los números correspondientes cumplen con que 1 es menor que 2 y resulta más sencillo escribirlo como 1<2.

Una variable está medida a nivel **ordinal** si los números que representan cada categoría son asignados de manera que respeten el orden según aumenta la característica que la variable mide. Estos números designan las categorías y son expresión de la jerarquía que hay entre ellas.

A continuación mostramos otros casos de variables medidas a nivel ordinal y su correspondiente codificación numérica:

Tabla 8: Ejemplo de variables de nivel ordinal

Código	Condición de pobreza
1	no pobre
2	pobre
3	indigente

Código	Rendimiento de los alumnos de una escuela
1	insatisfactorio
2	satisfactorio
3	bueno
4	muy bueno
5	excelente

De aquí en adelante ya no usaremos una columna especial de la tabla para indicar el código, simplemente lo señalamos junto al nombre de la categoría, como en las tablas siguientes:

Tabla 8 (continuación): Ejemplo de variables de nivel ordinal

Grado de participación de los padres en las reuniones convocadas por la escuela
0 nulo
1 bajo
2 moderado
3 alto

Año en el que se matriculó
1 Primero
2 Segundo
3 Tercero
4 Cuarto
5 Quinto

Acerca del significado de los valores numéricos en las variables de nivel ordinal, si bien hemos agregado el orden, aun no es posible hacer operaciones con ellos. Es decir, no es posible sumar dos valores y que la suma tenga algún significado. Por ejemplo, en la última variable de la tabla 8, no es cierto que 3=2+1, porque no es cierto que tercer año sea la suma de primero y segundo. Tampoco es válido restarlos, veamos que la diferencia entre 1 y 2 es 1 y la diferencia entre 3 y 4 también es 1, pero eso no tiene un correlato entre las categorías: no es cierto que haya la misma distancia entre primero y segundo año que entre tercero y cuarto, simplemente porque no tenemos definida la idea de distancia. ¿Qué podría significar la distancia entre cursos? ¿La dificultad para pasar de un año al siguiente?

Si 1 y 2 son dos categorías de una variable medida a nivel ordinal, se pueden establecer las relaciones: $1 \neq 2$ y $1 < 2$.

Los dos niveles (o escalas) de medición que consideraremos a continuación se llaman intervalares y proporcionales y usan las codificaciones numéricas con un significado un poco diferente al visto hasta aquí. La principal diferencia es que el grado de arbitrariedad que tenemos para asignar los números en variables medidas a estos niveles se reduce sustancialmente. Digamos de manera introductoria que en las escalas intervalares se conservan las distancias entre los valores: aquello que observamos que no podía hacerse en las ordinales, porque pasar de primer año a segundo no "es lo mismo" que pasar de tercero a cuarto. En las variables medidas a nivel proporcional, además de conservarse la distancia, se verifica la proporcionalidad de los valores: es decir que, recién en estas escalas, cuatro será el doble de dos.

El nivel intervalar

Veamos un ejemplo antes de definir este nivel. Cuando decimos que estamos en el año 2011, hacemos implícitamente una

afirmación que supone una medición del tiempo transcurrido desde un determinado evento, cuya elección no es única. En cierto modo decimos "han transcurrido 2011 años desde el momento que acordamos usar como inicio de este calendario". En culturas no cristianas, el origen en la medición de los tiempos puede ubicarse en otro momento y, en consecuencia, el año actual es otro. En el calendario judío, por ejemplo, el presente es el año 5771. Hay entonces cierto grado de arbitrariedad en la ubicación del punto desde donde empezar a contar los años. Lo que llamaríamos el "año cero", no es necesariamente el mismo. Sin embargo, el tiempo transcurrido entre 1975 y 2005 es de treinta años, como lo es el tiempo transcurrido entre 5735 y 5765. Es decir que la transformación de la que estamos hablando aquí, **conserva las distancias**. Independientemente de la escala con que hayamos medido el año, la diferencia entre dos años, se mantiene constante. Eso sucede porque las dos escalas (en este ejemplo, la medición del tiempo según las tradiciones cristiana y judía) se distinguen solo en la elección del origen (la posición del cero) pero no en la definición de lo que es un año. Para ambas escalas un año corresponde a una vuelta de la tierra al sol, por lo que la unidad de medición es la misma[11]. Ubicar el cero en un momento (en un determinado hecho histórico) o en otro es una elección; ese cero no indica la "ausencia de tiempo". En este caso, cero no quiere decir "nada", sino "origen elegido".
Llevemos esto a un terreno más cercano a la Psicología: al principio del capítulo dijimos que una forma de medir la inteligencia es la de observar cuántos problemas de una serie de dificultad creciente es cada uno capaz de resolver correctamente. Pero, ¿podríamos decir que quien no resuelve ninguno de ellos tiene inteligencia cero?, esto es claramente incorrecto, porque la ubicación del cero no implica la ausencia de lo que estamos midiendo (ausencia de inteligencia en este caso).

Las escalas intervalares, mantienen las propiedades de las escalas ordinales y nominales, es decir, los números designan categorías y permiten ordenarlas; pero además permiten decir a qué distancia está una de otra, porque cada categoría se expresa también en sentido cuantitativo. La medición intervalar implica construir una escala en la que las categorías están proporcionalmente distanciadas entre sí. Esto permite especificar la distancia que separa a cada categoría de las

[11] Si bien la corrección que se introduce cada año no es idéntica, por lo que el momento de cambio de año no es el mismo en las dos escalas.

demás. Este nivel de medición requiere que se establezca algún tipo de unidad de medida que pueda ser considerado por todos como una norma común y que sea repetible, esto es, que se pueda aplicar reiteradamente a los mismos individuos produciendo los mismos resultados. En el campo de la Psicología, especialmente en el uso de las pruebas mentales, como la medición de aptitudes, el uso de las escalas intervalares es muy frecuente.

Por ejemplo, la medición de los rendimientos individuales por medio de pruebas suele expresarse en puntajes que pueden provenir del tiempo requerido para realizar una determinada tarea o de la cantidad de trabajo realizado. En este tipo de prueba, es común que los puntajes partan de un mínimo establecido (por ejemplo el mínimo tiempo posible de ejecución o la mínima cantidad de tareas que una persona puede realizar en una prueba) y esto constituye el puntaje mínimo o la categoría más baja. Los puntajes de las pruebas mentales varían de acuerdo con el rendimiento y un mayor rendimiento siempre significará un mayor puntaje. En la medición de la inteligencia, es posible tomar los puntajes obtenidos en la prueba y categorizar, por ejemplo:

Tabla 9: Ejemplo de variable medida a nivel intervalar

Codificación numérica	Puntaje en una prueba	Significado
1	menos de 70	retraso significativo
2	de 70 a 85	retraso leve
3	de 85 a 100	normal
4	100 a 115	normal superior
5	más de 115	excepcional

Esta escala, con valores numéricos del 1 al 5, conservan las distancias, es decir que la distancia entre la categoría 1 y 2 es la misma que la que hay entre la 2 y la 3. Además de saber que un sujeto al que le corresponde el valor 4 tiene mayor inteligencia que uno al que le corresponde el valor 3 (información que ya nos pueden dar las escalas ordinales), sabemos también que la diferencia que hay entre esos dos individuos es de una unidad, y que es la mitad de la distancia que separa al que obtuvo 4 del que obtuvo 2. En esta escala las distancias entre observaciones son proporcionales.

Como lo indicamos antes, no hay un cero que pueda considerarse como la ausencia de aquello que se mide.

A este nivel de medición, ya es posible expresar las relaciones de manera formal; así, si x e y representan la medición del mismo atributo en diferentes escalas, puede obtenerse y a partir de x a través de la siguiente operación:

$$y = b_0 + b_1 * x$$

En la que b_0 y b_1 son números fijos elegidos arbitrariamente. El primero de ellos indica el desplazamiento en el origen de la escala: allí donde x valga 0, y tomará el valor de b_0. Por su parte, b_1 es un factor de escala, que modifica el tamaño de la unidad de medida. Veamos algunas aplicaciones.
El primer ejemplo fue el de la medición del año según dos calendarios diferentes, si llamamos x a la medición en el calendario cristiano e y a esa medición hecha con el calendario judío, tendremos:

$$y = 3760 + x$$

En la que hemos reemplazado b_0 por el valor 3760 y b_1 ha desaparecido, es decir que vale *1* (que no tiene efecto cuando multiplica a *x*). 3760 representa el cambio en el origen: cuando el calendario cristiano marcó cero (hipotéticamente, claro, porque su implementación es posterior a esa época), el judío indicaba el año 3760. El *1* correspondiente a b_1, e indica que no hay cambio en el tamaño de la unidad, como dijimos antes, ambas culturas acuerdan en que el año es una vuelta de la tierra al sol.

Una variable está medida a nivel **intervalar** cuando las distancias entre las categorías son proporcionales.

Si 1, 2, 3 y 4 son categorías de una variable medida a nivel intervalar, se pueden establecer las relaciones:

$$1 \neq 2$$
$$1 < 2$$
$$2 - 1 = 4 - 3$$

El nivel proporcional[12]
Este es el último nivel de medición que trataremos y se trata de aquél que más intuitivo nos resulta, es el único nivel considerado efectivamente como medición por la teoría clásica, ya que en él se integran todas las propiedades que hemos mencionado en los niveles anteriores y además se agrega la proporcionalidad de los valores numéricos y el carácter absoluto del cero. Podríamos decir que recién a este nivel, los números se

[12] Este nivel de medición aparece mencionado en alguna bibliografía como "escalas de razón" se pueden tratar como sinónimos, ya que la razón se refiere al cociente de números, que permanece constante en el caso de valores proporcionales.

comportan realmente como números, ya que se puede operar con ellos del modo al que estamos acostumbrados (sumarlos, multiplicarlos, etc.). ¿Qué variables pueden medirse a este nivel? Todas aquellas para las cuales tengan sentido las dos propiedades adicionales que esta escala incorpora: proporcionalidad de valores y cero absoluto. La cantidad de errores ortográficos cometidos en una prueba de dictado, admite el valor cero como correspondiente a "no errores", a la ausencia de lo que se mide, se trata de un cero absoluto. Además, cometer 10 errores es el doble que cometer 5. Por eso, la variable *Número de errores ortográficos cometidos* es de nivel proporcional. El *tiempo que una persona tarda en resolver una tarea*, si se mide en minutos, admite considerar que 4 minutos es el doble de 2, por lo que estamos también en presencia de una escala proporcional, aunque el cero no sea un valor observable. También es proporcional la variable *ingresos mensuales del hogar* o el *número de materias aprobadas*.
En general, los valores que provengan de procesos de conteo (como el número de errores) serán siempre proporcionales, como lo serán aquellas que hagan referencia a una unidad de medida estándar como el tiempo[13] o la distancia.

Tabla 10: Ejemplos de variables de nivel proporcional

Número de materias aprobadas (como regular) por alumnos que cursaron primer año
0
1
2
3
4
5
6

Cantidad de aplazos a lo largo de la carrera
0
1
2
3
4
5
6
7
8
9 ó más

[13] El ejemplo de los calendarios judío y cristiano, aunque es una medición de tiempo, no es absoluta. Es diferente de la medición con un cronómetro, que establece un inicio de cuenta al momento en que se lo dispara y da lugar a una variable de nivel proporcional.

Tiempo de reacción ante un estímulo visual (en segundos)
Menos de 5
Desde 5 hasta menos de 6
Desde 6 hasta menos de 7
Desde 7 hasta menos de 8
Desde 8 hasta menos de 9
Desde 9 hasta menos de 10
Más de 10

Así, cero significa ninguna materia aprobada o ningún aplazo. Tener cuatro materias aprobadas es el doble que tener dos, como 6 aplazos son el triple de 2. Estas variables respetan que el cero es absoluto (indica ausencia de lo que se mide) y que los valores son proporcionales.

En estos ejemplos hay algunas particularidades. El primero considera que seis es el número máximo de materias aprobadas como regular por quienes cursaron primer año, con lo que la categorización es exhaustiva, no es posible aprobar como regular más de seis materias en primer año. La segunda necesita dejar una categoría abierta final, porque el número de aplazos puede ser mayor a 10 y no resulta conveniente enumerar todos los posibles valores, porque la tabla sería muy molesta para leer. Observemos que esto es equivalente al recurso que usamos al referirnos al requisito de exhaustividad que debían cumplir las variables desde su menor nivel de medición. En el caso de las nominales, habíamos señalado que a veces es necesario incluir la categoría "otro" para asegurar la exhaustividad.

El tercer ejemplo tiene un problema parecido: los valores pueden ser más grandes que el límite superior y resolvemos este problema del mismo modo, con una categoría abierta final. Pero además, en esta variable son posibles los valores intermedios, alguien puede tardar 2,3 segundos para reaccionar ante el estímulo, ó 2,15 s y no es posible hacer una lista con todos los tiempos posibles. Este problema se resuelve construyendo intervalos, es decir, agrupando valores, por ejemplo desde 5 hasta menos de 6, etc.

En el próximo capítulo nos detendremos en las formas de construir estos agrupamientos; por ahora indiquemos solamente que los dos primeros ejemplos corresponden a variables que se llaman discretas, que quiere decir que solo pueden tomar valores enteros. Por el contrario, el tiempo de reacción es una variable continua, porque puede tomar todos los valores dentro de un intervalo, es decir puede cambiar gradualmente, no "salta" de un número entero al siguiente, como lo hace el número de materias aprobadas o la cantidad de aplazos.

Para la variable tiempo de reacción, es más frecuente presentar las categorías directamente así:

Tabla 11: Ejemplo de categorización de una variable de nivel proporcional continua

Tiempo de reacción ante un estímulo visual (en segundos)
Menos de 5
5-6
6-7
7-8
8-9
9-10
Más de 10

En la que es necesario establecer una convención sobre el intervalo al que pertenece cada valor discreto. Así, según la categorización anterior, el valor 6 pertenece al intervalo 6-7 y no al 5-6. Del mismo modo 7 pertenece al intervalo 7-8 y no al 6-7. De manera general, cada intervalo incluye al valor mínimo y excluye al máximo, que pasa a pertenecer al intervalo siguiente. Esto se debe a que hemos categorizado diciendo "desde… hasta menos de…".

Como en el nivel intervalar, expresemos estas propiedades de manera formal. Sean nuevamente x e y la medición del mismo atributo en diferentes escalas, ahora podemos obtener y a partir de x a través de la siguiente operación:

$$y = b_1 * x$$

En la que ahora solo tenemos un número fijo elegido arbitrariamente: b_1, que es el factor de escala, que modifica el tamaño de la unidad de medida. Esto simplemente significa que podemos cambiar las unidades con que medimos variables proporcionales: por ejemplo pasar de metros a centímetros, de horas a minutos, etc. Ninguna de esas transformaciones pueden modificar la posición del cero, porque en esta escala es absoluto: allí donde x valga cero, y deberá también valer cero, por eso no hay b_0.

Una variable está medida a nivel **proporcional** cuando sus valores respetan relaciones de proporcionalidad y, en consecuencia, el cero tiene un valor absoluto.

Si 1, 2, 3 y 4 son categorías de una variable medida a nivel proporcional, se pueden establecer las relaciones:

$$1 \neq 2$$
$$1 < 2$$
$$2 - 1 = 4 - 3$$
$$4 = 2 * 2$$

Cuadro 1: Resumen de las características de los diferentes niveles de medición

Nivel de medición	Significado de los símbolos numéricos	Requisito para cambiar los números	Ubicación del cero
Nominal	Designan, distinguen	Que no se repita el mismo para diferentes categorías	Sin significado
Ordinal	Expresan orden	Que respeten el orden de las categorías	Sin significado
Intervalar	Reflejan proporcionalidad de las distancias	$y = b_0 + b_1 * x$	Arbitrario
Proporcional	Reflejan proporcionalidad de los valores de la variable	$y = b_1 * x$	Absoluto (indica ausencia de lo que se mide)

Cuadro 2: Ejemplos de variables medidas a diferente nivel

	Nominales		Cepa de la que provienen los animales de laboratorio
	Ordinales		Grado de dificultad de un examen de ingreso a la Universidad
Métricas	Intervalares		Edad mental
	Proporcionales	Discretas	Número de palabras recordadas en una prueba de memoria
		Continuas	Duración de cada período de amamantamiento

Algunos elementos teóricos de la discusión sobre medición

En función de lo expuesto, estamos en condiciones de referir brevemente algunos debates clásicos sobre este tema. Según Galtung (1968) todo dato hace referencia a una estructura constituida por tres elementos: unidad de análisis, variable y valor; así cualquier dato aislado sería: "una unidad de análisis que en una variable específica presentará un determinado valor". Esto significa que estos tres elementos deben ser considerados conjuntamente para sostener una proposición empírica. Así un sistema tiene diferentes propiedades que pueden ser observadas y aportarán diferentes datos; por ejemplo en el campo de la Psicología, un sistema sería la personalidad de los sujetos observados (unidades de análisis), pero según el aspecto observado tendremos distintas propiedades (variables) derivadas del mismo sistema y que nos permitirían distintos tipos de medición; a cada sujeto le corresponderá una categoría de cada una de esas variables (valor). Supongamos que dentro de la personalidad estamos midiendo rasgos de apatía, una variable a observar podría ser el tiempo que demora una persona en decodificar una determinada orden mientras que otra variable podría ser la actividad que prefiere realizar en sus ratos libres. Puede notarse que los números asignados a estas propiedades significarán cuestiones muy distintas en uno u otro caso: la primera variable se medirá en unidades de tiempo (segundos, por ejemplo) y será de nivel proporcional; la segunda tendrá categorías como "hacer deportes", "leer", "ir al cine", etc., por lo que tendrá nivel de medición nominal.

Galtung señala que: "dado un conjunto de unidades, un valor es algo que puede predicarse de una unidad y una variable es un conjunto de valores que forma una clasificación". Ese conjunto de valores no pueden ser menos de dos, es decir la variable debe poder "variar" entre por lo menos dos valores que la conforman.

Sobre la escala ordinal, Selltiz (1980) indica que: "la escala ordinal define la posición relativa de objetos o individuos con respecto a una característica, sin implicación alguna en cuanto a la distancia entre posiciones".

Según Garret (1974), el orden determina rangos de los objetos de estudio, pero dichos rangos sólo indican una posición serial en el grupo, sin darnos una medida exacta; no podemos sumar o restar rangos como si fueran centímetros o kilómetros: el rango de una persona o hecho observado siempre es relativo en comparación con los rangos de los otros elementos observados y jamás en términos de alguna unidad conocida.

Acerca de la diferencia entre escalas intervalares y proporcionales Blalock (1966) señala que "esta distinción... es

puramente académica ya que es extremadamente difícil encontrar una escala legítima de intervalos que no sea al propio tiempo una escala de proporciones. Esto se debe al hecho de que, una vez establecida la magnitud de la unidad, casi siempre es posible concebir cero unidades... Así pues, prácticamente en todos los casos en que se dispone de una unidad, será legítimo emplear todas las operaciones matemáticas".

Actividad práctica de repaso 1

Considere las variables:
Edad, Cantidad de materias aprobadas, Promedio a lo largo de la carrera, Método anticonceptivo usado, Carrera que cursa

Y las siguientes, con sus categorías respectivas:

Puntaje en la escala de inteligencia de Wechsler	Depresión (Escala de Beck)
130 o más	Normal
120-129	Ligero trastorno emocional
110-119	Depresión clínica borderline
90-109	Depresión moderada
80-89	Depresión grave
70-79	
50-69	
49-30	

Título máximo alcanzado por docentes de una Facultad
Doctorado
Maestría
Licenciatura
Tecnicatura

Aunque no hay valores numéricos, es posible decidir el nivel de medición de cada variable si se tienen en cuenta las propiedades que cumplen las categorías. Por ejemplo, género admite como categorías varón y mujer, que solo se distinguen entre ellas y no pueden ordenarse. Por lo tanto, género es una variable nominal.

1. Indique el nivel de medición de cada una de las variables mencionadas.

2. Identifique los juicios válidos en las variables:

	$1 \neq 2$	$1 < 2$	$3-1 = 4-2$	$1 = \frac{1}{2} * 2$
Edad				
Cantidad de materias aprobadas				
Método anticonceptivo usado				
Promedio a lo largo de la carrera				
Carrera que cursa				
Nivel de depresión				
Título máximo de los docentes				
Puntaje en la escala de inteligencia de Wechsler				

3. Defina categorías para las siguientes variables e indique el nivel de medición de cada una:

Cantidad de materias rendidas desde que ingresó a la carrera
Síntomas somáticos propios de la ansiedad
Razones para la consulta a un hospital neuropsiquiátrico
Actitud hacia la participación política
Concepto que los docentes tienen de los alumnos.

Capítulo 2: La organización de los datos

Eduardo Bologna

En este capítulo veremos procedimientos que sirven para presentar la información de manera accesible para que pueda ser interpretada. Veremos que para poder extraer significado de los datos recogidos es necesario primero dedicar un esfuerzo a organizarlos, a presentarlos de manera comprensible.

De la información en bruto a la matriz de datos

El primer paso en la descripción de un conjunto de datos es el de organizar la información recogida con la construcción de la llamada matriz de datos. Supongamos que hemos administrado una encuesta a 150 personas y que el siguiente es un fragmento del cuestionario usado.

Cuestionario No: .....

Sexo ☐ 1 masculino ☐ 2 femenino

Edad ..... años

Actualmente usted es: ☐ 1 soltero ☐ 2 casado o unido ☐ 3 separado o divorciado ☐ 4 viudo

¿Qué hizo durante la mayor parte del tiempo la semana pasada? ☐ 1 trabajó ☐ 2 buscó trabajo ☐ 3 estudió ☐ 4 realizó otra actividad

¿Cuáles fueron los ingresos de su hogar el mes pasado? ..... pesos

¿Cuántas personas habitan esta vivienda? ..... Personas

Una vez completados los cuestionarios, la información está "en bruto" y es necesario ordenarla para poder tener una visión de conjunto. Eso se logra organizando los datos recogidos en la **matriz de datos** que tiene, para el fragmento de cuestionario mostrado, la siguiente forma

orden	sexo	edad	estado civil	actividad la semana anterior	ingresos del hogar	Personas en la vivienda
1	1	25	1	3	2500	2
2	1	30	1	2	1650	3
3	2	21	2	2	720	1
4	2	57	2	3	3280	2
5	2	40	1	4	2700	2
…						
150	2	43	3	1	2000	4

Se trata de un ordenamiento de la información que contiene en la primera fila (horizontal) los nombres de las variables[14] y en las filas siguientes los números que corresponden a las respuestas dadas por los encuestados. Así, la persona que respondió al primer cuestionario es un varón (1) de 25 años, soltero (1) quien durante la semana pasada estudió, cuyos ingresos familiares ascienden a 2.500 pesos y que vive con una persona más en la vivienda (2 personas en total).

Mirando desde las variables: las frecuencias simples

Por su parte, cada columna (vertical) de la matriz de datos corresponde a una variable, esto es lo que nos permitirá ahora presentar la información de manera resumida. Si se lee la columna encabezada "sexo" pueden contarse cuántos unos (1s) y cuántos dos (2s) hay en total. En nuestra matriz de datos el recuento indica que, de los 150 casos, 78 son dos y 72 unos. Esto puede decirse brevemente así:

[14] Nótese que los nombres de las variables no son idénticos a las preguntas del cuestionario, a menudo la pregunta debe formularse de manera más comprensible para el encuestado. En todos los casos la forma de preguntar debe adaptarse al lenguaje del grupo al que se interrogará. En Metodología de la Investigación se verá este tema con más detalle.

Tabla 1	
sexo	casos
1	72
2	78
n	150

O de manera más explícita:

Tabla 2	
sexo	casos
masculino	72
femenino	78
n	150

Que nos informan, simplemente, que hay 72 varones y 78 mujeres. A la cantidad de casos, que proviene del recuento del número de 1s y 2s en la columna de sexo, se lo llama técnicamente **frecuencia absoluta simple** y se la indica como **f**. La tabla resulta entonces:

sexo	f
1	72
2	78
n	150

El total de 150 casos resulta de la suma de todas las frecuencias absolutas simples, de manera breve, esto se indica así:

$$\sum_{i=1}^{k} f_i = n$$

Que se lee "La sumatoria de las frecuencias desde *1* hasta *k* es igual al total de observaciones".
En esa expresión, Σ es el símbolo de suma o sumatoria e indica la realización de esa operación (sumar).
- Las f_i son las frecuencias absolutas simples. El subíndice *i* va cambiando entre categorías.
- La expresión *i=1* señala desde qué valor de *i* se inicia la suma, así como *k* señala la última categoría a sumar. En el ejemplo de la tabla 3, el valor de *k* es 2 (solo hay dos categorías), por lo que solo hay dos frecuencias a sumar: f_1 y f_2, correspondientes a varones y mujeres.
- *n* es el total de casos (observaciones).

Así, lo mismo puede indicarse como:

$$f_1 + f_2 + \cdots + f_k = n$$

Lo cual, en el caso de la tabla 3 resulta simplemente:

$$f_1 + f_2 = 72 + 78 = 150$$

La **frecuencia absoluta simple** de cada valor de la variable es el número de casos que asumen ese valor. Se indica *f*.

Si se quisieran comparar estas frecuencias con las de otra encuesta que tuviera un número total de casos diferente de 150, no sería útil usar los valores absolutos aquí presentados. Veamos por ejemplo si la comparación es entre la tabla que acabamos de mostrar y otra de la que solo sabemos que contiene 90 varones. La información disponible solo nos diría que en una muestra hay 72 varones y en la otra 90. Sobre esos números no podemos hacer ningún juicio, ya que para saber si son muchos o pocos, o si hay más o menos varones en una muestra o en la otra, necesitamos el total. Si bien 90 es más que 72, la comparación depende de cuál sea el total *sobre* el que se lo cuente. Sabemos que la primera muestra tiene un total de 150 casos, si la otra tiene 200, podríamos afirmar que la cantidad de varones es parecida (poco menos de la mitad en ambos casos). Pero para comparar con certeza nos hace falta indicar el peso *relativo* de los varones, no su número total, sino su contribución al total de casos.
Calcularlo es muy sencillo, ya que solo debemos dividir el número de varones en el total general. En nuestro ejemplo 72/150 es 0,48 que también puede leerse como 48%. Es decir que los varones constituyen una proporción de 0,48 o bien que representan el 48% del total. Esta proporción se denomina **frecuencia relativa simple**, se simboliza como *f'* (efe prima), y se calcula como acabamos de mostrar: dividiendo la frecuencia absoluta por el total. Ahora puede completarse la tabla anterior agregando otra columna.

Tabla 4

sexo	f	f'
1	72	0,48
2	78	0,52
n	150	1,00

El valor 1,00 que resulta de sumar las dos frecuencias relativas corresponde al 100% de los casos, es decir a las 150 observaciones. Usando la misma simbología que antes diremos ahora que:

$$\sum_{i=1}^{k} f'_i = 1$$

Que afirma que la suma de las frecuencias relativas simples (*f'*) es igual a uno.

Repitamos esta operación para la variable Situación conyugal[15]

Tabla 5

Situación conyugal	f	f'
1	45	0,3
2	30	0,2
3	15	0,1
4	60	0,4
n	150	1,00

La tabla dice que hay 30 personas casadas, que constituyen el 20% del total (*f'=0,2*), y del mismo modo con las demás categorías de la variable.

Observemos que al construir estas tablas de distribución de frecuencias hemos renunciado a una parte de la información que estaba en la matriz de datos. En ella podíamos seguir por la fila a cada individuo como hicimos con la primera y describirlo en cada una de sus variables. Por el contrario, la tabla de distribución de frecuencias solo nos dice que hay 72 varones y 78 mujeres o que hay 45 solteros y 30 casados, pero no nos dice quiénes son. Esta pérdida de información es parte inevitable del proceso en el que resumimos los datos, cuanto más sintética sea la presentación, tanta más información habremos perdido. Esto puede visualizarse como el proceso en el que vamos "tomando distancia" de los datos originales: cada vez tenemos una mejor visión de conjunto, pero al mismo tiempo perdemos detalles.

La **frecuencia relativa simple** de cada valor de la variable es la proporción de casos que asumen ese valor. Se indica *f'*.

Los dos ejemplos mostrados hasta aquí corresponden a variables medidas a nivel nominal, por lo que los números no son más que códigos, no representan orden ni puede considerarse la distancia entre ellos. ¿Qué cambia si trabajamos con un nivel de medición más elevado? Veamos lo que sucede con la última variable de la matriz del ejemplo, el número de personas que habitan la vivienda. Con el mismo principio que usamos para las variables nominales, la forma de la tabla de distribución de frecuencia sería:

[15] Recordemos que sus categorías fueron definidas como:1, soltero; 2, casado o unido; 3, separado o divorciado; 4, viudo

Tabla 6

Número de personas en la vivienda	f	f'
1	4	0,03
2	10	0,07
3	13	0,09
4	41	0,28
5	24	0,16
6	26	0,17
7	31	0,21
n	150	1,00

La variable que se considera aquí está medida a escala proporcional y los valores que puede asumir solo son enteros, porque no hay fracciones de personas para contabilizar. Sin embargo, a fin de generalizar la construcción de categorías, haremos "como si" cada una de ellas fuera un intervalo, que comienza media unidad por debajo y termina media unidad por encima de cada valor real. Así, el valor 1 corresponde al intervalo 0,5 - 1,5. Esto equivale a decir que en lugar de contar 1 persona decimos que hay entre 0,5 y 1,5, pero como ese número debe ser entero, solo puede ser el 1. Aunque parezca muy artificioso usar esta notación, lo hacemos porque nos trae la ventaja de poder usarse de la misma manera para todas las variables de nivel proporcional, lo que será especialmente valioso en aquellas que sí admiten valores fraccionarios. Obsérvese que al hacer esta transformación hemos pasado de unas categorías que "saltaban" de un valor entero al siguiente (de 1 a 2, de 2 a 3, etc.) a otras que ahora cambian gradualmente: allí donde termina la primera categoría (en 1,5) se inicia la siguiente. Diremos que la variable se presenta en categorías (o **clases**) de una unidad y también que la **amplitud** de cada clase (o categoría) es de una unidad. La tabla no cambia su aspecto, pero ahora diremos que los **límites exactos** son los que se encuentran media unidad por debajo y media unidad por encima de cada valor. Los números 0,5 y 1,5 son los límites exactos de la primera categoría (o clase), 1,5 y 2,5 son los límites exactos de la segunda y así sucesivamente.

Cuando las categorías de una variable se agrupan en intervalos (clases), se indica también el punto medio de cada intervalo, llamado **marca de clase** (MC), que se calcula promediando los límites inferior y superior de cada clase.

El paquete InfoStat® presenta esta tabla del siguiente modo:

Tabla 7

Tablas de frecuencias

Variable	Clase	LI	LS	MC	FA	FR
personas en la vivie	1	0,50	1,50	1,00	4	0,03
personas en la vivie	2	1,50	2,50	2,00	10	0,07
personas en la vivie	3	2,50	3,50	3,00	13	0,09
personas en la vivie	4	3,50	4,50	4,00	41	0,28
personas en la vivie	5	4,50	5,50	5,00	24	0,16
personas en la vivie	6	5,50	6,50	6,00	26	0,17
personas en la vivie	7	6,50	7,50	7,00	31	0,21

La primera columna indica el nombre de la variable, repetido en cada categoría y truncado a veinte caracteres; la columna encabezada "clase" es el número de intervalo (no son valores de la variable), luego se muestran los límites exactos inferior y superior (LI y LS) de cada clase, la marca de clase (MC), la frecuencia absoluta (FA) y frecuencia relativa (FR).

Al observar el caso de la variable *Edad*, en nuestra matriz de datos, nos encontramos con un problema adicional, ya que el número de categorías que admite es bastante más elevado que las que venimos viendo hasta acá; es decir, al cuestionario pueden haber respondido personas de las edades más variadas y no resultaría cómodo mostrar en una lista todos y cada uno de los valores, digamos, entre 18 y 75 años, si esas son, por ejemplo, las edades menor y mayor de las persona encuestadas. Para ello será necesario construir categorías de mayor amplitud que una unidad, es decir, categorías que agrupen a varios valores. La tabla siguiente es un ejemplo de esa agrupación:

Tabla 8

Edad (agrupada)	f	f'
18-27	59	0,39
28-37	55	0,37
38-47	14	0,09
48-57	10	0,07
58-67	7	0,05
68-77	5	0,03
n	150	1,00

En este ejemplo hemos decidido construir seis categorías (o clases). Los límites exactos de la primera categoría son 17,5 y 27,5; los de la segunda son 27,5 y 37,5 y así sucesivamente. Resulta entonces que la variable está agrupada en clases de 10

unidades, la amplitud de cada clase es de 10 años, que resulta de la diferencia (resta) entre el límite exacto superior y el inferior de cualesquiera de ellas (27,5-17,5=10; 37,5-27,5=10; etc.). De manera simbólica:

$$A = L_s - L_i$$

Expresión en la que A es la amplitud de la clase, L_s es el límite exacto superior y L_i, el límite exacto inferior de esa clase.
¿Cómo se decide la amplitud de cada clase? Se trata de un compromiso entre la simplicidad de la presentación que se logra con pocas categorías y la pérdida de información que implica hacer clases de mucha amplitud. Por regla general, cuantas menos clases se construyan (y, en consecuencia, de mayor amplitud) mayor será la pérdida de información, pero un exceso de categorías dará lugar a una presentación poco clara.

No hay diferencias en esta presentación si la variable admite categorías con decimales, como, en nuestro ejemplo, los ingresos mensuales del hogar. Si elegimos clases de $1000, la siguiente es la salida InfoStat® correspondiente

Tabla 9

Tablas de frecuencias

Variable	Clase	LI	LS	MC	FA	FR
ingreso	1	0,00	1000,00	500,00	112	0,21
ingreso	2	1000,00	2000,00	1500,00	180	0,34
ingreso	3	2000,00	3000,00	2500,00	190	0,36
ingreso	4	3000,00	4000,00	3500,00	50	0,09

Aquí se ve más claro que la "clase" indica el número de categoría y no el valor de la variable. El número de categorías —que como vimos depende del grado de detalle con que quieran mostrarse los datos—, puede ser elegido de manera automática por InfoStat®, o bien cambiarse manualmente. Por ejemplo, podemos pedir más clases, para dar más detalles sobre la distribución del ingreso, con intervalos de $500:

Tabla 10

Tablas de frecuencias

Variable	Clase	LI	LS	MC	FA	FR
ingreso	1	0,00	500,00	250,00	49	0,09
ingreso	2	500,00	1000,00	750,00	63	0,12
ingreso	3	1000,00	1500,00	1250,00	75	0,14
ingreso	4	1500,00	2000,00	1750,00	105	0,20
ingreso	5	2000,00	2500,00	2250,00	120	0,23
ingreso	6	2500,00	3000,00	2750,00	70	0,13
ingreso	7	3000,00	3500,00	3250,00	50	0,09

O bien podemos pedir menos clases, para ofrecer solo una visión general, a "grandes rasgos", con pocas y amplias categorías:

Tabla 11

Tablas de frecuencias

Variable	Clase	LI	LS	MC	FA	FR
ingreso	1	0,00	1500,00	750,00	187	0,35
ingreso	2	1500,00	3000,00	2250,00	295	0,55
ingreso	3	3000,00	4500,00	3750,00	50	0,09

Observemos que en estos ejemplos los límites exactos de las clases son los que se indican en las tablas, esto depende del programa que se utilice. Si las clases están separadas, entonces los límites exactos están media unidad por debajo y por encima de los que aparecen en la tabla. Si las clases son contiguas, los límites exactos son los que aparecen en la tabla.

De acuerdo a estos ejemplos, nos encontramos con dos situaciones en que apelamos a la presentación de los valores de la variable en forma de intervalos: si se trata de una variable discreta con muchas categorías (como la edad) o si es una variable continua.

Variable discreta con muchas categorías
La construcción de intervalos es una elección; podríamos optar por mostrar todas las categorías, con lo que quedaría una tabla grande, pero muy detallada; o bien agrupar para ganar en sencillez de presentación. Es muy común optar por la construcción de intervalos, de manera de mantener la cantidad de categorías entre cinco y diez. En tablas en que se precisa mostrar mucho detalle, se opta por la enumeración de todas las categorías.

Variable continua
Si la variable es continua no podemos elegir, porque no se pueden mostrar "todas las categorías" de una variable continua, ya que éstas son infinitas. Para verlo consideremos que en un conjunto continuo, entre dos números siempre puede hallarse otro, por ejemplo: entre 2,50 y 2,60 está el 2,55; entre una persona con ingresos mensuales de 2576,20 y otra que tiene 2576,30, podría haber alguien con 2576,25. Es por esto que los valores no cambian en cantidades fijas como los discretos, sino de manera continua, gradual. Consideremos el caso de la estatura: en términos matemáticos es correcto decir que no hay persona alguna que mida exactamente 1,75 m, porque "exactamente" quiere decir 1,750000…. (y siguen los ceros), que equivale a afirmar que sería posible medir la estatura con precisión infinita, a fin de asegurar que la décima, centésima, milésima, etc., de metro son todos ceros. Como esto no puede hacerse, porque nuestros instrumentos de medición tienen precisión finita, nunca sabremos si estamos en presencia de alguien que mida exactamente 1,75m. Esta idea

puede parecer un tanto abstracta, lo que importa recordar de ella es que, para variables continuas, no es posible indicar frecuencias simples de valores individuales, solo de intervalos de valores. En el ejemplo, diremos que la frecuencia (absoluta o relativa) simple de 1,75m es siempre cero, como lo es la de cualquier valor único. Por el contrario, no hay inconveniente en indicar cuántas personas (o qué proporción de ellas) tiene estatura entre 1,749 y 1,751, si tenemos una regla que mida con precisión de un milímetro.
Las frecuencias que pueden indicarse para variables continuas pueden corresponder a un intervalo entre dos valores —como en el ejemplo—, o bien a partir de un valor dado hacia los valores mayores o hacia los menores. Podemos responder a ¿cuántas personas (o qué proporción de ellas) tiene más de 1,75m o menos de 1,75m?

La siguiente tabla muestra nuevamente la distribución del ingreso, en siete intervalos, con las frecuencias absolutas y relativas, con el formato en que lo presenta InfoStat®.

LI	LS	FA	FR
100	943	16546	0,58
943	1786	7114	0,25
1786	2629	2787	0,10
2629	3471	1035	0,04
3471	4314	535	0,02
4314	5157	241	0,01
5157	6000	148	0,01

Medidas usuales relacionadas con las frecuencias relativas
Las tasas, razones y proporciones, son también frecuencias relativas, solo que por el uso se les ha dado nombre específico.
Así, nos referimos habitualmente (pero no siempre) como **tasa** a la frecuencia relativa de un fenómeno en referencia a una población total, con la característica de tener en cuenta un período de tiempo. También es común el uso del término cuando se trata de hechos de poca incidencia, es decir que su frecuencia es pequeña. En esos casos se la suele expresar cada 1.000, cada 10.000, o inclusive cada 100.000 casos. Por ejemplo:

Defunciones [anuales] por accidentes de vehículos de motor. México 1972[16]

Grupos de edad	Tasa de mortalidad (por 100.000)
0	1,8
1-4	3,4
5-14	4,1
15-24	10,1
25-34	12,6
35-44	12,0
45-54	15,3
55-64	17,6
65-74	22,5
75 ó más	24,7

La lectura de esta tabla debe hacerse considerando que los totales sobre los que se calculan corresponden al total de personas de cada grupo de edad. Así, la segunda de las tasas indica que se produjeron 3,4 muertes de personas entre 1 y 4 años por cada 100.000 personas de esas edades.
Las tasas así calculadas se llaman tasas específicas, en este caso, por grupos de edades y se distinguen de las tasas brutas (o crudas). Por ejemplo, la tasa bruta de mortalidad indica el cociente entre el total de defunciones ocurridas en un año y la población total (estimada para la mitad del año).

La palabra **razones** se usa a menudo para referirse a cocientes calculados entre conjuntos que no tienen elementos en común. Por ejemplo, se llama razón de masculinidad a la cantidad de hombres por cada 100 mujeres que hay en una población. Se obtiene dividiendo el total de varones por el total de mujeres (y luego multiplicando por 100), que son dos conjuntos que no se superponen. Esta medida se conoce también como índice de masculinidad.
Por ejemplo: el grupo de estudiantes que cursó Estadística en Psicología en 2010 estuvo compuesto por 1050 mujeres y 280 varones, por lo que la razón de masculinidad es de 27 (que resulta de 280/1050*100), al que leemos como 27 varones por cada 100 mujeres.
En la población general de la ciudad de Córdoba, según el censo de 2010, la razón (o índice) de masculinidad es de 91,9 varones cada 100 mujeres.

[16] Tomado de Guerrero et al (1986). Material bibliográfico de la Cátedra de Psicología Sanitaria.

Usamos **proporción** para indicar el cociente entre una parte de la población y el total. Por ejemplo, la proporción de personas de 65 años afectadas de Alzheimer es aproximadamente 1,5 cada 100, lo que suele expresarse con forma de porcentaje para facilitar la lectura: 1,5%. Esto indica que, del total de personas de 65 años, ese porcentaje está afectado de la enfermedad.

Las frecuencias acumuladas

En los últimos párrafos nos hemos concentrado en el modo en que se presentan las categorías de la variable y dejamos de mencionar a las frecuencias por un momento. Recordemos que la frecuencia absoluta indica la cantidad de observaciones en cada categoría, o el número de casos a quienes corresponde ese valor de la variable y que la frecuencia relativa indica la proporción de casos en cada categoría, que puede leerse (si se multiplica por 100) como el porcentaje de casos en cada categoría. Así, en la tabla 9 (la de los ingresos en cuatro categorías) diremos que 180 personas tienen ingresos del hogar entre $1000 y $2000, mientras que la proporción de quienes los tienen entre $2000 y $3000 es de 0,36 ó, lo que es lo mismo que el 36% tiene ingresos entre $2000 y $3000.

Además de indicar cuántos casos (o qué porcentaje de ellos) tiene determinados valores de la variable, resulta de interés mostrar cuantos (y también que porcentaje) tienen valores *iguales o menores* a uno determinado. Por ejemplo, en la distribución de la tabla 9, además de saber cuántos tienen ingresos entre $2000 y $3000, también interesa saber cuántos tienen menos de $3000. Igualmente sucede con las proporciones o los porcentajes. Esta información solo tiene sentido si se refiere a variables que respeten el orden de sus categorías, puede preguntarse por “menos de 2000 pesos”, “menos de tres personas”, “menos de 45 años” pero no puede interrogarse por “menos de soltero”. Esto último no tiene sentido porque la variable situación conyugal no tiene sus categorías ordenadas, ya que su nivel de medición es nominal.

Para responder a la pregunta por la cantidad de casos que hay por debajo de una categoría de la variable (solo para variables medidas a escala ordinal o superior) usaremos las **frecuencias acumuladas**. Su cálculo es muy simple, ya que solo es necesario contar las frecuencias de la categoría que nos interesa y sumarla a las frecuencias de las categorías anteriores a ella. Volvamos sobre el ejemplo de las edades

Tabla 12

Edad (agrupada)	f	f'	F
18-27	20	0,13	20
28-37	40	0,27	60
38-47	20	0,13	80
48-57	10	0,07	90
58-67	40	0,27	130
68-77	20	0,13	150
n	150	1,00	

Hemos agregado otra columna a la tabla anterior a la que rotulamos con *F*, que contiene las **frecuencias absolutas acumuladas**, las que resultan de la operación que recién mencionamos: la primera categoría tiene frecuencia acumulada igual a la absoluta simple, porque no hay ningún caso por debajo de 17,5 (este es el límite exacto inferior de la primera clase); la segunda es 60, que proviene de contar los 40 de la segunda categoría y sumarle los 20 de la anterior. Del mismo modo se construyen las siguientes. La última categoría tiene por frecuencia absoluta acumulada al total de casos (en el ejemplo 150), porque todos (los 150) están en esa categoría o por debajo de ella. La lectura que hacemos de estas frecuencias es que, por ejemplo, "hay 80 personas que tienen 47 años[17] o menos." ¿Por qué 47? Porque la frecuencia acumulada reune los casos *de esa categoría* y las anteriores, por lo que los 20 que están entre 38 y 47 están también allí contados.

La **frecuencia absoluta acumulada** de cada valor de la variable es la cantidad de casos que asumen ese valor y todos los valores menores a él. Se indica *F*.

Por la misma razón que expusimos antes —la necesidad de comparar distribuciones de frecuencia que tengan diferente número de casos—, es necesario disponer de valores *relativos* de estas frecuencias, que se llamarán **frecuencias acumuladas relativas**. La regla para obtenerlas es la misma que para las frecuencias simples: se divide cada frecuencia absoluta acumulada en el total de casos. Al agregar esta frecuencia, la tabla anterior resulta:

[17] Estrictamente deberíamos decir "que tienen 47,5 años o menos", ya que ese es el límite exacto superior de esa clase. Para simplificar la lectura lo expresamos como 47.

Tabla 13

Edad (agrupada)	f	f'	F	F'
18-27	20	0,13	20	0,13
28-37	40	0,27	60	0,40
38-47	20	0,13	80	0,53
48-57	10	0,07	90	0,60
58-67	40	0,27	130	0,87
68-77	20	0,13	150	1,00
n	150	1,00		

La lectura de una de estas frecuencias es, por ejemplo, para la tercera categoría: "la proporción de quienes tienen 47 años o menos es de 0,53", a los fines de la comunicación puede ser más sencillo presentarlo como porcentaje: "el 53% de los individuos tiene 47 años o menos".

La **frecuencia relativa acumulada** de cada valor de la variable es la proporción de casos que asumen ese valor y todos los valores menores a él. Se indica F´.

Cuando se solicitan a InfoStat®, estas salidas tienen la forma siguiente (para el ejemplo del número de personas en la vivienda):

Tabla 14

Tablas de frecuencias

Variable	Clase	LI	LS	MC	FA	FR	FAA	FRA
personas en la vivie	1	0,50	1,50	1,00	4	0,03	4	0,03
personas en la vivie	2	1,50	2,50	2,00	10	0,07	14	0,09
personas en la vivie	3	2,50	3,50	3,00	13	0,09	27	0,18
personas en la vivie	4	3,50	4,50	4,00	41	0,28	68	0,46
personas en la vivie	5	4,50	5,50	5,00	24	0,16	92	0,62
personas en la vivie	6	5,50	6,50	6,00	26	0,17	118	0,79
personas en la vivie	7	6, 50	7,50	7,00	31	0,21	149	1,00

Tabla en la que se han agregado las columnas FAA y FRA correspondientes a frecuencia absoluta acumulada y frecuencia relativa acumulada respectivamente.

Las frecuencias acumuladas tienen especial interés para las variables continuas, dado que —como mencionamos más arriba— en ellas no pueden indicarse las frecuencias simples de un valor. Sí en cambio será posible indicar la frecuencia acumulada hasta ese valor. No podremos responder a la pregunta "¿cuántos miden exactamente 1,75m?", pero sí podemos usar la frecuencia acumulada para responder a

"¿cuántos miden 1,75m o menos?". Éste es el tipo de pregunta que podemos responder sobre variables continuas.
Vemos esto en el siguiente ejemplo, con la variable (continua) tiempos de reacción a un estímulo auditivo, medida sobre una muestra de 34 sujetos experimentales:

Tiempo de reacción (en décimas de segundo, ds)	f	f´	F	F´
1,0-1,5	5	0,15	5	0,15
1,5-2,0	7	**0,21**	12	0,35
2,0-2,5	6	0,18	**18**	0,53
2,5-3,0	**3**	0,09	21	0,62
3,0-3,5	8	0,24	29	**0,85**
3,5-4,0	5	0,15	34	1,00
n	34	1,00		

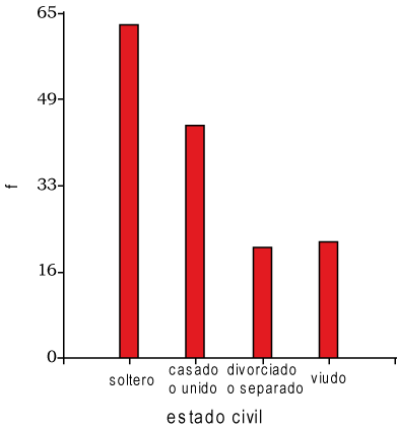
Acerca de los valores destacados, leemos de esta tabla que:
Tres personas mostraron tiempos de reacción entre 2,5 y 3,0 ds (frecuencia absoluta simple).
El 21% (una proporción de 0,21) de los sujetos experimentales tuvo tiempos de reacción entre 1,5 y 2,0 ds (frecuencia relativa simple).
18 sujetos tuvieron tiempos de reacción por debajo de 2,5 ds (frecuencia absoluta acumulada).
El 85% (una proporción de 0,85) de los sujetos tuvo tiempos de reacción por debajo de 3,5 ds (frecuencia relativa acumulada).

¿Cómo presentar de manera gráfica los resultados?
En la misma dirección de ofrecer una presentación de los datos recogidos que sea accesible para la interpretación, veremos a continuación las representaciones gráficas más frecuentemente usadas para mostrar información cuantitativa. Nuevamente aquí deberemos sacrificar la cantidad de información que se ofrece, a cambio del valioso impacto visual y facilidad de lectura que proveen los gráficos.
Cuando se trata de variables nominales, normalmente con pocas categorías, son adecuados los **gráficos de barras** o los **diagramas de sectores circulares** (o "de torta"). Veamos un ejemplo para la tabla de la situación conyugal que reproducimos a continuación:

Tabla 15

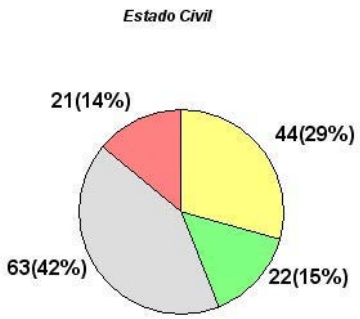
Situación conyugal	f	f'
soltero	63	0,42
casado o unido	44	0,29
divorciado o separado	21	0,14
viudo	22	0,15
total	150	1,00

El paquete InfoStat® presenta el gráfico de barras así:

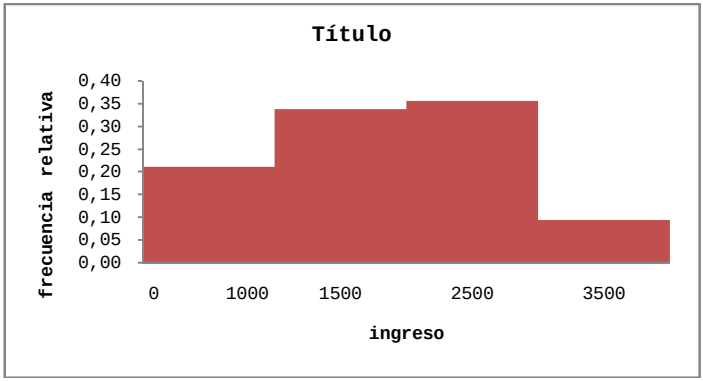


Y del siguiente modo los gráficos de sectores:

Al que resulta posible modificar en cuanto a formato, rótulos, etc., por ejemplo, si solicitamos que muestre las frecuencias absolutas y relativas de cada categoría, resulta:



En los casos en que la variable tiene categorías cuantitativas (intervalar o proporcional) se utiliza un gráfico llamado **histograma**. Este gráfico no debe confundirse con el de barras, que se usa con variables nominales.
La presentación de InfoStat® para el ejemplo de los ingresos de la tabla 9 es:

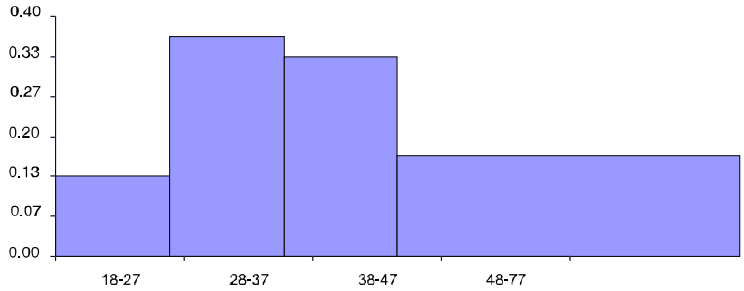


El lugar donde dice "título" es un espacio editable, para escribir el título del gráfico que elijamos.
A pesar de su simplicidad hay un aspecto a tener en cuenta en este gráfico, que será importante más adelante. Es el hecho que se trata de un gráfico de áreas, ¿qué quiere decir esto? Veamos un ejemplo un poco diferente, supongamos que las amplitudes de las clases no son iguales, que, por ejemplo, hay muy pocos casos en las categorías más altas y que decidimos agrupar juntos a todos los que tienen más de 47 años. La tabla quedaría ahora así:

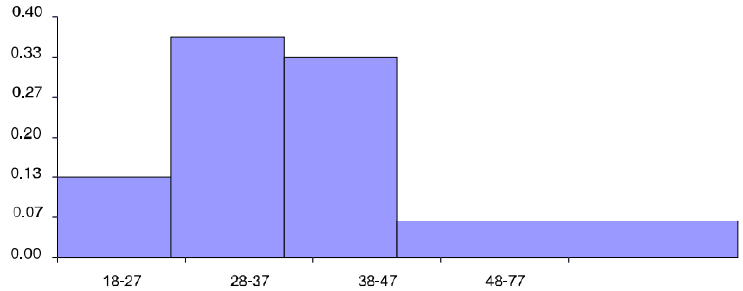
Tabla 16

Edad (agrupada)	f	f'
18-27	20	0,13
28-37	55	0,37
38-47	50	0,33
48-77	25	0,17
total	150	1,00

Si graficamos sin tener en cuenta la agrupación, el gráfico tendrá la forma siguiente:



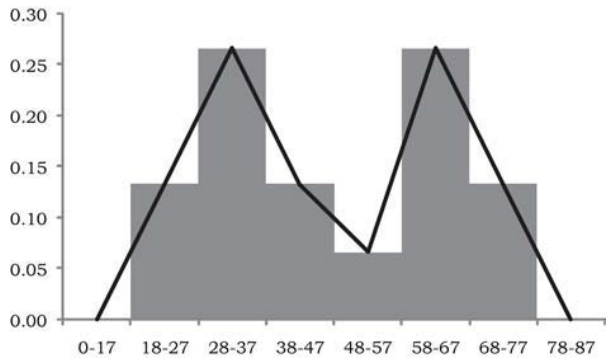
En esta representación la categoría 48-77 aparece como muy importante, y esto no sucede porque tenga mucha frecuencia sino porque es más ancha (tiene mayor amplitud); aun así, el efecto visual confunde, porque hace creer que son edades de mayor importancia que la real. Para eliminar este efecto indeseable, se calcula la altura correspondiente a la frecuencia considerando que es la *superficie* y no la altura la que la representa, y se obtiene:



Así, las clases que sean más amplias tendrán menor altura que la que les correspondería por su frecuencia (para que la *superficie del rectángulo = base por altura*, sea proporcional a la frecuencia).

No es importante saber hacer esa cuenta, pero sí es muy importante recordar que el histograma es un gráfico de superficie: es el área (o superficie) de las barras y no su altura la que indica la frecuencia. En consecuencia, la suma de las superficies de todas las barras será igual al total de casos (*n*) si graficamos frecuencias absolutas, y dará uno (*1*) si las que se grafican son las relativas.

Los histogramas pueden transformarse en **polígonos de frecuencias** uniendo los puntos medios de cada intervalo como se muestra a continuación (volvemos al ejemplo de clases de igual amplitud, con los datos de la tabla 13).

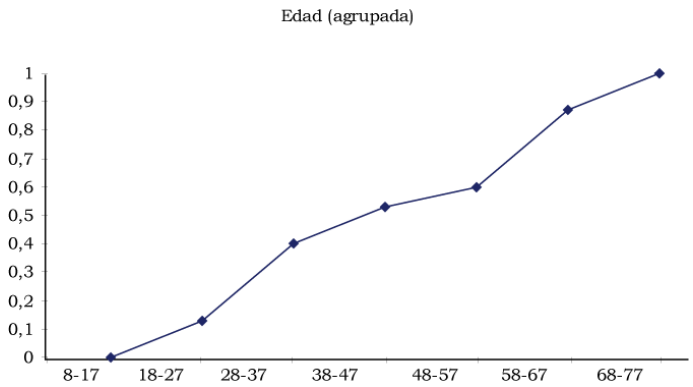


En este gráfico hemos agregado dos intervalos, uno anterior al primero y uno posterior al último, cuyas frecuencias son cero, con el objetivo de "cerrar" el polígono sobre el eje horizontal.
El área que queda bajo este polígono es igual a la que encierran los rectángulos del histograma, y valdrá *n* si se grafican frecuencias absolutas ó *1* si son las relativas, como en este ejemplo y como más comúnmente se hace, ya que permite comparar distribuciones de frecuencia que tengan diferente número de casos.

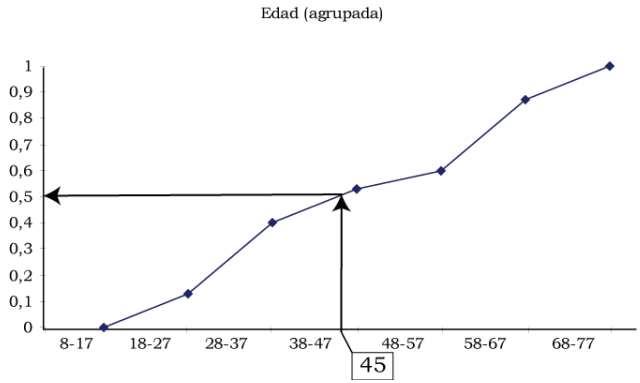
Como ya señalamos, en este tipo de variables (intervalares o proporcionales) es posible calcular frecuencias acumuladas, por lo que también ellas pueden representarse gráficamente.

Tabla 17

Edad (agrupada)	f	f'	F	F'
18-27	20	0,13	20	0,13
28-37	40	0,27	60	0,40
38-47	20	0,13	80	0,53
48-57	10	0,07	90	0,60
58-67	40	0,27	130	0,87
68-77	20	0,13	150	1,00
total	150	1,00		



Este gráfico se llama **ojiva**.
Obsérvese que la frecuencia acumulada para cada categoría se representa con un punto que corresponde al límite superior de cada una, esto es por la misma razón de antes: lo acumulado hasta esa categoría la incluye a ella.
Hemos agregado una categoría más, correspondiente a las edades 8 a 17 años; es la anterior a la primera que aparece en la tabla. A esta categoría le corresponde frecuencia acumulada igual a cero y la incluimos para cerrar el gráfico sobre el eje horizontal.
Este gráfico tiene otra virtud además de la claridad visual, ya que permite **interpolar** valores no observados, o que no aparecen en la tabla. Así, con el gráfico podemos responder a la pregunta ¿Qué proporción de casos tiene 45 años o menos? Como el valor 45 años no aparece en la tabla sino dentro de una categoría, no es posible responder desde la tabla; sin embargo, en el gráfico podemos buscar el valor 45 años e identificar la frecuencia acumulada que le corresponde.



En este ejemplo, la ordenada (valor en el eje vertical) correspondiente a los 45 años es aproximadamente 0,5 (0,50), este resultado se lee diciendo que el 50% de los encuestados tienen 45 años o menos. En los capítulos siguientes veremos otras aplicaciones útiles de este procedimiento.

Resumen de definiciones presentadas en el capítulo

Frecuencia	Símbolo	Significado	Nivel en que tiene interpretación
Absoluta simple	f	Cantidad de observaciones en cada categoría de la variable	Todos
Relativa simple	f'	Proporción de observaciones en cada categoría de la variable	Todos
Absoluta acumulada	F	Cantidad de observaciones en cada categoría de la variable y en todas las anteriores a ella	Ordinal o superior
Relativa acumulada	F'	Proporción de observaciones en cada categoría de la variable y en todas las anteriores a ella	Ordinal o superior

Actividad práctica de repaso 2

En un trabajo desarrollado por el Laboratorio de Psicología Cognitiva de la Facultad de Psicología, se llevó a cabo una investigación dentro del marco de la teoría Psicolingüística. En este estudio dirigido por Manoiloff y Seguí, se realizó un relevamiento de datos en una muestra de 35 estudiantes universitarios con el objeto de evaluar a qué edad se adquirían determinadas palabras. Algunos de los resultados obtenidos fueron:

"Cangrejo"

Edad de Adquisición	Frecuencia	Porcentaje
0 a 3 años	1	2,9
3 a 6 años	17	48,6
6 a 9 años	17	48,6
Total	35	100

"Manzana"

Edad de Adquisición	Frecuencia	Porcentaje
0 a 3 años	16	45,7
3 a 6 años	14	40,0
6 a 9 años	2	5,7
9 a 12 años	3	8,6
Total	35	100

"Binoculares"

Edad de Adquisición	Frecuencia	Porcentaje
3 a 6 años	2	5,7
6 a 9 años	10	28,6
9 a 12 años	10	28,6
más de 12 años	13	37,1
Total	35	100

1. Responda a las siguientes preguntas:
a. ¿Cuántas personas de la muestra adquirieron la palabra "cangrejo" entre los 3 y los 6 años?
b. ¿A qué edad se adquiere con mayor frecuencia la palabra "binoculares"?
c. ¿Cuál es la palabra que se adquiere más temprano según los estudios reportados?

Dadas las siguientes tablas de distribución de frecuencias

Coeficiente intelectual				
Superior	20	20	0,03	0,03
Brillante	50	**70**	0,10	0,07
Inteligente	70	140	0,20	0,10
Normal	150	290	**0,41**	0,21
Poco inteligente	180	470	0,67	0,26
Limítrofe (borderline o fronterizo)	100	570	0,81	**0,14**
Deficiencia mental superficial	60	630	0,90	0,09
Deficiencia mental media	**40**	670	0,96	0,06
Deficiencia mental profunda	30	700	1,00	0,04
Total	700			1,00

Tipo de delito				
Robo	150	**150**	0,28	0,28
Lesiones leves	240	90	**0,17**	0,45
Hurto	440	200	0,37	**0,82**
Lesiones graves	**520**	80	0,15	0,97
Asesinato	535	15	0,03	1,00
Total		535	1,00	

Cantidad de materias aprobadas				
0	200	0,10	200	0,10
1	**450**	0,13	250	0,23
2	750	0,15	**300**	0,38
3	1050	0,15	300	0,53
4	1450	0,20	400	**0,74**
5	1750	0,15	300	0,89
6	1950	**0,10**	200	0,99
7	1970	0,01	20	1,00
Total		1,00	1970	

2. Para cada una:

a. Indique el nivel de medición de cada variable.
b. Rotule las columnas según se trate de frecuencias absolutas o relativas, simples o acumuladas. (Atención a que en las tablas aparecen desordenadas).
c. Señale qué frecuencias tienen significado según el nivel de medición de las variables.
d. Redacte una interpretación para cada uno de los valores que se encuentran destacados en las tablas y que tengan significado.

Capítulo 3: La expresión resumida de la información

Eduardo Bologna

La segunda etapa en la descripción de un conjunto de datos consistirá en calcular medidas que los resuman, que los expresen de manera sintética. Esta etapa implicará un nuevo alejamiento de la información bruta, ya que perderemos de vista no solo a los individuos —que aparecían en la matriz de datos—, sino también a las distribuciones de frecuencia. La ventaja de los procedimientos que veremos en este capítulo es la posibilidad de presentar la información de modo muy sintético; con unas pocas medidas descriptivas ofreceremos bastante información sobre los datos que se han recogido.

Digamos antes de empezar que estas medidas requieren operaciones de diferente nivel de complejidad, por lo que apelan a diferentes propiedades de las escalas de medición, entonces no serán las mismas las medidas que se puedan calcular en una escala nominal que en una ordinal, intervalar o proporcional.

El objetivo de describir el conjunto de datos se logrará indicando tres tipos diferentes de medidas. En primer lugar, haremos referencia a las medidas de **posición**. Estas medidas nos indicarán en torno a qué valores se distribuyen las observaciones. Dentro de las medidas de posición, definiremos las medidas centrales, (también llamadas de **centralidad** o de **tendencia central**), y no centrales. En segundo lugar, mencionaremos las medidas de **dispersión** (conocidas también como de **variabilidad**), que mostrarán si los datos están concentrados alrededor de las medidas de centralidad o si están dispersos, alejados de esas medidas centrales. En tercer lugar, nos detendremos en la forma que asume la distribución y allí, aunque hay otras medidas, solo nos ocuparemos de describir la **simetría** o asimetría que manifiesta el conjunto de datos.

A los fines de la notación usada para referirse a cada una de estas medidas descriptivas, asumiremos que trabajamos sobre datos provenientes de una muestra, de la que n representa la cantidad de casos observados.

Medidas de posición

Entre las medidas que resumen una distribución de frecuencias, mencionaremos las centrales y las no centrales. Las medidas que se puedan calcular dependerán del nivel de medición de las variables que se describan, por lo que las presentaremos separadamente para cada nivel, siempre recordando que las operaciones que son válidas a un determinado nivel de medición también son válidas para niveles más altos. Por ejemplo: lo que pueda hacerse con variables nominales, vale también para ordinales y métricas.

Medidas de centralidad

Son las que indican alrededor de qué valores de ubican las observaciones de una distribución de frecuencias.

Variables nominales: la proporción

Cuando se trabaja con una variable de nivel nominal, una manera sintética de presentar la información que ofrece la tabla de distribución de frecuencias es indicando la **proporción** de casos que se encuentran en una determinada categoría. Se trata de la frecuencia relativa simple (f') de una categoría particular. Sea la siguiente una clasificación de los diagnósticos dados por un psicólogo a un conjunto de pacientes:

Tabla 1

Diagnóstico	f	f'
Psicosis	10	0,125
Neurosis	50	0,625
Perversión	20	0,250
Total	80	1,000

Podemos indicar la proporción de casos diagnosticados como psicosis, como *p=0,125*, que puede también expresarse como 12,5%. La elección de cuál categoría se elige para indicar la proporción solo depende de los objetivos de la descripción. Al elegir una categoría se llama la atención sobre ella, se la destaca, ya que la proporción restante incluye a todas las demás categorías, los "otros". Esa proporción restante se obtiene restando de 1 (uno) la proporción indicada, o restando de 100 (cien) si ha expresado como porcentaje. En nuestro ejemplo, diremos que 0,875 (que proviene de hacer 1-0,125) es la proporción de *otros diagnósticos* o bien que éstos representan el 87,5% (100-12,5).

La **proporción** es la frecuencia relativa correspondiente a una categoría particular. Puede expresarse como decimal o en porcentaje. Se indica como *p*.

Esta medida descriptiva se usa a menudo cuando la variable nominal tiene solo dos categorías, ya que se presenta la proporción de una de ellas e inmediatamente se sabe que el complemento es la proporción de la otra. Si se trata de pacientes que consultan a un servicio de admisión psicológica en un hospital, ellos pueden ser clasificados como sigue:

Tabla 2

Resultado de la entrevista de admisión	f	f̃
Admitido como paciente	150	0,75
No admitido	50	0,25
Total	200	1,00

Eligiendo como categoría de referencia "Admitido como paciente", resulta ser *p=0,75*, que dice que la proporción de pacientes admitidos es de 0,75, o del 75%. Si se resume la tabla diciendo que "el 75% de las personas que consultan es admitido", se sabe de inmediato que el 25% restante no es admitido.

Notemos que esta medida es la misma que presentamos en el capítulo 2 cuando indicamos que la proporción es el cociente entre la frecuencia propia de la categoría y el total de casos. Esta proporción puede también indicarse en variables de nivel de medición superior al nominal, pero no resulta de interés cuando hay gran cantidad de categorías. Así, por ejemplo, si se trata de la distribución de las notas de un parcial, no se estila indicar cuál es la proporción de cada calificación (lo que se vería en una tabla de distribución de frecuencias de las notas). Sin embargo, es común construir variables nominales a partir de las notas y es de mucho interés indicar, por ejemplo, la proporción de *promocionados*, o la proporción de quienes *quedaron libres*.

Variables nominales: el modo

La más elemental de las medidas de centralidad que se usa en los distintos niveles de medición se denomina **modo**, o **moda**, o **valor modal** y es simplemente el valor de la variable (la categoría) que tiene la mayor frecuencia. Dicho de otra manera, el valor de la

variable más frecuentemente observado[18]. Esta medida no requiere ningún cálculo, no exige ninguna propiedad de la escala de medición, por lo tanto se puede indicar en variables desde el nivel nominal, es decir en todos los niveles de medición.

La variable *tipo de hogar*, tiene la siguiente distribución:

Tabla 3

tipo de hogar	f
unipersonal	40
nuclear	90
extendido	20
ampliado	10
total	160

El modo es hogar de tipo "nuclear", que es la categoría de mayor frecuencia. Debe cuidarse de no cometer el error de señalar la frecuencia 90 como el modo; el modo no es la frecuencia más alta, sino la categoría de la variable que tiene mayor frecuencia. Para hallarlo, se identifica la más alta de las frecuencias y se señala la categoría que le corresponde.

Si se trata de una variable de mayor nivel de medición, no hay ninguna diferencia. La variable *concepto que los docentes asignan a los alumnos* tiene la distribución de frecuencias siguiente:

Tabla 4

concepto	f
Excelente	150
Muy bueno	350
Bueno	200
Satisfactorio	120
No satisfactorio	50
Total	870

En este ejemplo, 350 es la frecuencia más alta, por lo tanto, la categoría que a ella corresponde es el modo: el modo de la distribución es "Muy bueno".

[18] Esta es la idea que transmite el lenguaje coloquial: cuando algo es "la moda", es lo que más comúnmente (frecuentemente) se ve.

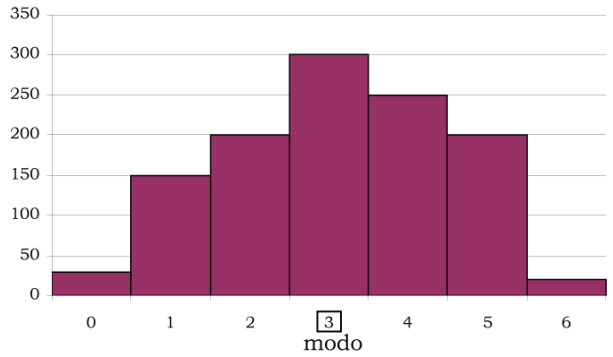
El **modo** es la categoría —o el valor— de la variable que tiene mayor frecuencia. Se indica M_o

Cuando se trabaja sobre variables intervalares o proporcionales discretas no hay diferencia en la identificación del modo de la distribución. El *número de materias que tienen aprobadas alumnos que han terminado de cursar el primer año* de su carrera se distribuye así:

Tabla 5

Número de materias aprobadas	f
0	30
1	150
2	200
3	300
4	250
5	200
6	20
Total	1150

En esta distribución, el modo es 3 materias aprobadas ($M_o=3$), que es la categoría que tiene mayor frecuencia. Expresamos esto como "la mayor cantidad de alumnos que terminaron de cursar primer año han aprobado tres materias". Si se observa el histograma correspondiente a esta distribución, el modo aparece claramente en la categoría que tiene la mayor superficie (en este caso se trata simplemente de la mayor altura de los rectángulos pero, como vimos en el capítulo 2, si las amplitudes fueran diferentes debe considerarse la superficie de los rectángulos como representativa de la frecuencia).

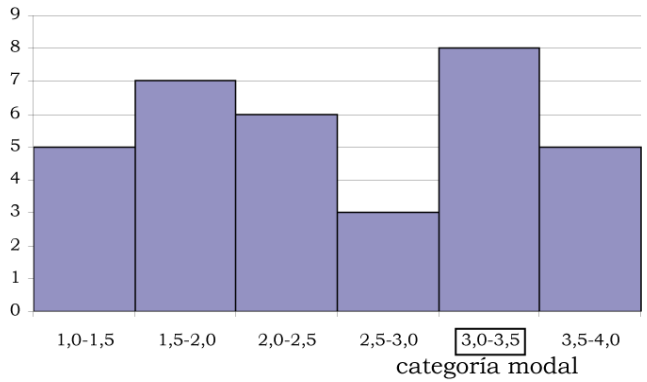


Cuando se trata de datos agrupados en clases de más de una unidad, ya no resulta posible hallar un valor único que sea el modo, sino una categoría, que en este caso es un intervalo. Por ejemplo, si tratamos con la variable *tiempo de reacción ante un estímulo auditivo*.

Tabla 6

Tiempo de reacción (en segundos)	f (sujetos experimentales)
1,0-1,5	5
1,5-2,0	7
2,0-2,5	6
2,5-3,0	3
3,0-3,5	8
3,5-4,0	5
Total	34

La mayor frecuencia se encuentra en la categoría 3,0-3,5 y ésa es la categoría modal. No hay en este caso un valor único del tiempo que se llame "el modo", sino una categoría de máxima frecuencia a la que llamamos modal. De manera equivalente a la variable anterior, el gráfico permite una identificación inmediata de esta categoría:



Para identificar un valor único dentro de la categoría modal, se realiza una interpolación a su interior, para lo que se usa la siguiente expresión:

$$M_o = L_i + i * \left(\frac{f_{post}}{f_{ant} + f_{post}} \right)$$

En la que L_i es el límite inferior de la categoría modal, i es la amplitud de esa categoría, f_{ant} es la frecuencia absoluta de la categoría anterior y f_{post} es la frecuencia absoluta de la categoría

posterior. Para usarla se debe primero identificar a la categoría modal y luego a las frecuencias anterior y posterior. Aplicada a los datos de la tabla 6 resulta:

$$M_o = 3{,}0 + 0{,}5 * \left(\frac{5}{3+5}\right) = 3{,}0 + 0{,}5 * 0{,}625 = 3{,}0 + 0{,}312 = 3{,}312$$

Observemos el orden en que se hacen las operaciones: la suma separa términos, por lo que primero se resuelve el paréntesis, se multiplica por la amplitud y recién entonces se suma el límite inferior.

Puede suceder que en una distribución no haya una única categoría de mayor frecuencia, sino que dos o más compartan la mayor frecuencia. Para 160 alumnos clasificados según la facultad en que cursan su carrera, tenemos:

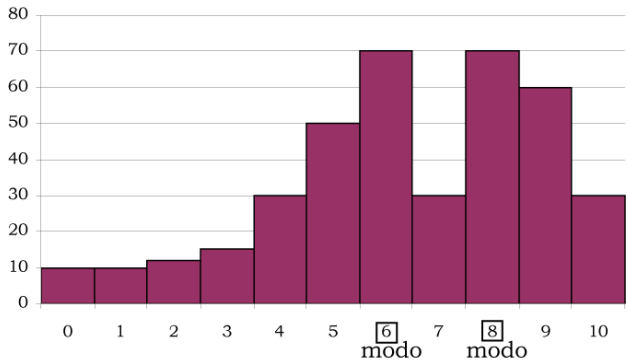
Tabla 7

Facultad a la que pertenece	f
Arquitectura	50
Ingeniería	40
Psicología	50
Filosofía	20
Total	160

Vemos aquí que hay dos categorías que presentan la mayor frecuencia: Arquitectura y Psicología. Decimos en este caso que la distribución es **bimodal** que quiere decir simplemente que tiene dos modos.

Una distribución es **bimodal** cuando dos categorías tienen la mayor frecuencia. Si son más las categorías que comparten la mayor frecuencia, la distribución se denomina multimodal

Una representación gráfica de una distribución bimodal, para la variable *número de respuestas correctas en una prueba de opción múltiple*, la siguiente



La moda tiene el inconveniente de ser independiente de la mayor parte de los datos, por lo que es sensible a cambios en los valores de la distribución. En efecto, las siguientes dos muestras de 130 escuelas tienen la misma moda (M_o=*Pública*), aunque son muy dispares:

Gestión de la escuela	f
Pública	50
Privada laica	45
Privada confesional	35
total	130

Gestión de la escuela	f
Pública	100
Privada laica	20
Privada confesional	10
total	130

Además, en tablas con categorías de más de una unidad, su valor cambia según cuántas categorías se construyan.

Variables de nivel ordinal: la mediana

Como ya hemos visto, cuando las categorías de la variable están ordenadas pueden hacerse juicios como "mayor que" (>) o "menor que" (<), y el nivel de medición es ordinal. En este tipo de variables puede calcularse otra medida de centralidad, que usa esa propiedad: la del orden entre categorías. Se trata de la **mediana**, que se podrá también calcular para escalas superiores (intervalar y proporcional) pero no para escalas nominales, en las que el orden entre categorías no está presente. Esta medida se define como el valor de la variable que deja por debajo la mitad del total de observaciones. Se trata de la mitad de los casos y no la mitad de las categorías. La siguiente distribución presenta en serie simple el número de sesiones de psicoterapia que recibieron 9 pacientes internados en un hospital:

2, 2, 3, 5, 7, 10, 15, 19, 50

La mediana de estos datos es 7, porque es el valor que deja cuatro casos por debajo y también cuatro casos por encima. Atención a que la serie simple debe estar ordenada para poder identificar a la mediana.

Si la cantidad de observaciones fuera par como por ejemplo:

2, 2, 3, 5, 7, 10, 15, 19

El punto de corte correspondiente a la mitad de las observaciones se ubica entre 5 y 7, en este caso, la mediana es el promedio entre los dos valores centrales, que es 6.

Cuando hay valores repetidos en la parte central no resulta posible indicar la mediana, por ejemplo si la serie fuera:

2, 3, 7, 7, 7, 10, 15, 19

No puede señalarse a 7 como la mediana, porque es superado por tres valores (10, 15 y 19) pero supera solo a dos (2 y 3).

La mediana es una medida muy adecuada cuando se necesitan resumir datos que provienen de escalas ordinales o de nivel superior. Sin embargo, su cálculo no es muy usual en series simples con pocos casos (como los que acabamos de ver) ya que allí es más sencillo mostrar el conjunto completo de datos y no usar medidas resumen.

Se denomina **mediana** al valor de la variable que deja por debajo a la mitad de las observaciones. La mediana deja la misma cantidad de casos por debajo y por encima de ella. Se indica M_{dn}.

Veamos la forma de reconocer a la mediana cuando los datos están presentados en una distribución de frecuencias. La siguiente es una clasificación de hogares por nivel socioeconómico:

Tabla 8

Nivel socioeconómico	f (número de hogares)	F
Marginal	40	40
Bajo	100	140
Medio-bajo	120	260
Medio	150	410
Alto	30	440
Total	440	

La mitad del número total de casos es 220, que resulta de dividir por dos los 440 casos del total (440/2, equivalente a (1/2)*440). Entonces, para hallar la mediana, se debería identificar al caso (al hogar en el ejemplo) que ocupa el lugar 220. Según la tabla, hasta la categoría "medio-bajo" se acumulan 260 hogares y hasta la categoría anterior hay 140 acumulados. De manera que el hogar que ocupa el lugar 220 es uno de los que se encuentran en la categoría "medio-bajo". Diremos así que la mediana de esta distribución es "medio-bajo". Como puede verse, esta categoría no acumula exactamente la mitad de las observaciones, pero es la que contiene a la observación que supera a la mitad y es superada por la otra mitad. Por esa razón la leemos como la mediana, diciendo que la mitad de los hogares tiene un nivel socioeconómico medio-bajo o inferior a ese.

La imprecisión señalada en este ejemplo también aparece si se trata de una variable cuantitativa que se presenta en clases de una unidad (una variable discreta). Si la variable es el *número de síntomas* a partir de los cuales fueron diagnosticados de esquizofrenia un conjunto de pacientes[19]:

Tabla 9

Número de síntomas	f (número de pacientes)	F
2	30	30
3	60	90
4	70	160
5	90	250
Total	250	

La mitad del número total de casos es de 125 (250/2 ó (1/2)*250), ¿cuál es el caso que ocupa ese lugar? Vemos que hasta 3 síntomas se acumulan 90 pacientes y 160 hasta los 4. El paciente que ocupa el puesto 125 en la serie simple ordenada es uno de los que fueron diagnosticados a partir de 4 síntomas. Por esa razón indicamos a la mediana con ese valor: 4, $M_{dn}=4$. Nuevamente encontramos que no es exactamente el valor que acumula la mitad, sino uno de los que están en la categoría dentro de la cual se acumula la mitad de los casos. La primera

[19] El manual DSM IV indica como síntomas característicos a dos (o más) de los siguientes, cada uno de ellos presente durante una parte significativa de un período de 1 mes: ideas delirantes, alucinaciones, lenguaje desorganizado (p. ej., descarrilamiento frecuente o incoherencia), comportamiento catatónico o gravemente desorganizado, síntomas negativos, por ejemplo, aplanamiento afectivo, alogia o abulia.

categoría que tenga una frecuencia acumulada superior al 50% será la que contenga a la mediana. Es así porque la clase anterior no alcanza a acumular la mitad de las observaciones. La lectura del resultado es que la mitad de los pacientes fue diagnosticada de esquizofrenia en presencia de cuatro síntomas o menos.

Apliquemos el mismo procedimiento al *número de materias aprobadas*:

Tabla 10

Número de materias aprobadas	f	F
0	30	30
1	150	180
2	200	380
3	300	680
4	250	930
5	200	1130
6	20	1150
Total	1150	

La mitad de 1150 es 575, la primera frecuencia acumulada que supera a 575 es 680, que corresponde al valor 3. La mediana de esta distribución es entonces tres materias aprobadas y diremos que el 50% de los alumnos aprobó tres materias o menos.

Cuando se trata de variables de nivel intervalar o proporcional y con categorías agrupadas en clases de más de una unidad, el cálculo anterior puede refinarse. Así, primero identificaremos la categoría (la clase) en que se encuentra la mediana, de la forma que lo hicimos en los dos ejemplos anteriores y luego haremos una interpolación dentro del intervalo para encontrar su valor exacto. Esta interpolación es parecida a la que nos permitió calcular la moda.

Tabla 11

Tiempo de reacción (en segundos)	f	F
1,0-1,5	5	5
1,5-2,0	7	12
2,0-2,5	6	18
2,5-3,0	3	21
3,0-3,5	8	29
3,5-4,0	5	34
Total	34	

La mitad de las 34 observaciones es 17, por lo que debe encontrarse una observación que tenga frecuencia acumulada de 17. Ese valor no aparece en la F, el primero que lo supera es 18, entonces la mediana estará en el intervalo 2,0-2,5. Esto es así porque hasta 2,0 se acumulan 12 casos (la F de la categoría anterior) y hasta 2,5 se acumulan 18. Nuestros 17 casos se acumulan para un valor de la variable que está entre 2,0 y 2,5.
Debemos ahora encontrar qué valor exactamente es la mediana, dentro del intervalo 2,0-2,5. La fórmula para este procedimiento es la siguiente:

$$M_{dn} = l_i + i * \left(\frac{\frac{n}{2} - F}{f_p} \right)$$

En la que:
l_i indica el límite exacto inferior del intervalo en que se encuentra la mediana, en este caso es 80.
i es la amplitud del intervalo, es decir la diferencia entre los límites exactos 2,0-2,5 = 0,5
$\frac{n}{2}$ es la mitad del número total de observaciones, es este caso, 17
F_d es la frecuencia acumulada por **d**ebajo de la categoría que contiene la mediana, en esta tabla es 12
f_p es la frecuencia **p**ropia del intervalo en que se encuentra la mediana. Es la frecuencia absoluta (no la acumulada), en este ejemplo es 6.

Reemplazando resulta:

$$M_{dn} = 2{,}0 + 0{,}5 * \left(\frac{17 - 12}{6} \right) = 2{,}0 + 0{,}5 * \left(\frac{5}{6} \right) = 2{,}42$$

Conviene detenerse en el orden en que se realizaron las operaciones. El signo más (+) separa términos, por lo que debe primero resolverse el segundo de ellos y luego recién sumar 2,0. Un error frecuente es el de sumar 2,0 + 0,5 y luego multiplicar por el resultado del paréntesis, eso es incorrecto.
Por cierto que debe verificarse que el valor encontrado se ubique dentro del intervalo; en este ejemplo, la mediana no podría ser menor que 2,0 ni mayor que 2,5. Observemos también que el número de categorías de la variable, que es de 6, no participa en el cálculo de la mediana, de ningún modo se trata de una categoría que esté "al medio".

El resultado obtenido nos dice, según la definición de la mediana que "el 50% de los sujetos experimentales reaccionó en un tiempo de 2,42 segundos o inferior". Es muy importante la

última parte de la lectura, porque cuando decimos "o inferior" incluimos los valores por debajo del indicado.
La mediana encontrada, de 2,42, es un valor razonable a partir de la observación de la tabla: la categoría de la mediana acumulaba 18 casos, que es apenas más que la mitad de las observaciones (17), por lo que era de esperar que la mediana apareciera cerca del límite superior del intervalo, que es lo que sucedió.

Variable métricas: la media o promedio
Si se ha alcanzado un nivel de medición intervalar o proporcional, es posible hacer uso de las propiedades[20] que estas escalas tienen. Recordemos que además de designar y ordenar, las escalas intervalares conservan las distancias entre observaciones, y las proporcionales agregan la proporcionalidad de los valores absolutos. En este nivel los números que representan las categorías (o valores) pueden tratarse como tales y se puede operar con ellos. Antes de dar una definición de la media o promedio, veamos la idea intuitiva que tenemos, ya que se trata de una medida de mucho uso. Cuando queremos calcular un promedio "sumamos y dividimos por la cantidad de casos". Así, si tres personas cometen 5, 8 y 12 errores cada uno, el promedio de esa variable (número de errores) es $\frac{5+8+12}{3} = \frac{25}{3} = 8{,}33$. Usaremos la expresión $\bar{x}$ para referirnos a la media, con lo que $\bar{x} = 8{,}33$ errores.

¿Cómo extenderemos esta forma de cálculo al caso en que la variable no está presentada en serie simple sino en distribución de frecuencias? Recordando que la frecuencia indica la cantidad de veces que cada valor se repite, por lo que habrá que considerar cada valor tantas veces como lo indique su frecuencia absoluta simple. Veamos un ejemplo en el que se cuenta el número de materias aprobadas:

[20] Propiedades que se agregan a las de las escalas de menor nivel, por lo que modo y mediana pueden calcularse e interpretarse también en las escalas métricas.

Tabla 10

Número de materias aprobadas	f
0	30
1	150
2	200
3	300
4	250
5	200
6	20
Total	1150

El valor 0 (cero) está repetido 30 veces, lo que indica que hay 30 alumnos que no han aprobado aun ninguna materia. Del mismo modo, 150 alumnos aprobaron 1 materia, etc. para calcular el promedio de materias aprobadas por el conjunto de alumnos multiplicaremos cada valor de la variable por su frecuencia y dividiremos por el total de casos. Resulta:

$$\bar{x} = \frac{0*30+1*150+2*200+3*300+4*250+5*200+6*20}{1150} = 3{,}10$$

La expresión formal de este cálculo es:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i * f_i}{n}$$

En la que x_i es cada valor de la variable, f_i es su frecuencia absoluta simple, k es el número de categorías y n es el total de observaciones. La fórmula indica que cada valor de la variable (x_i) se multiplica por su frecuencia (f_i), se suman desde el primero ($i=1$) hasta el último (k) y el resultado se divide por el total de casos (n).

Veamos que no se trató como podríamos haber pensado rápidamente, de sumar desde el cero hasta el seis y dividir por siete. Haber hecho eso habría implicado dos errores: el primero es el de no considerar cuántas veces está repetido cada valor (su frecuencia absoluta simple), el segundo es el de confundir el número de casos (1150) con el número de categorías (7). Este último error puede provenir de una confusión entre la presentación en serie simple o en distribución de frecuencias. Cuando se observa una serie simple, los valores "sueltos" de la variable coinciden con sus categorías, pero cuando se agrupa, cada categoría incluye varios valores, que están indicados en la frecuencia de cada categoría[21].

[21] Si esto no resulta perfectamente claro, conviene releer el capítulo 2: La organización de los datos.

En el ejemplo anterior entonces, el número promedio de materias aprobadas es 3,10. Este número no es entero y no es un valor que se pueda observar; nadie tiene 3,10 materias aprobadas. Sin embargo, es valioso para caracterizar a la distribución completa y para hacer comparaciones. Por ejemplo, si en un grupo de alumnos la media es de 3,10 materias aprobadas y en otro de 3,90; puede decirse que en el segundo grupo los alumnos han aprobado —en promedio— más materias; aunque ninguno haya aprobado 3,10 ni 3,90 materias.
Por el momento ofreceremos una definición operacional de la media, más adelante en este capítulo podremos dar una definición conceptual, basada en sus propiedades.

La **media** (o promedio) es un valor de la variable obtenido sumando todas las observaciones multiplicadas por su frecuencia absoluta y dividiendo el resultado en el número total de casos. Se indica como $\bar{x}$ (equis media).

Cuando la distribución de frecuencias presenta los datos en clases de más de una unidad (datos agrupados), nos encontramos con el problema de no tener un único valor en cada categoría. Por ejemplo y nuevamente en el caso de los tiempos de reacción:

Tabla 11

Tiempo de reacción (en segundos)	f
1,0-1,5	5
1,5-2,0	7
2,0-2,5	6
2,5-3,0	3
3,0-3,5	8
3,5-4,0	5
Total	34

Vemos que no hay un valor único en cada categoría, sino un intervalo que incluye diferentes valores. Resolvemos este inconveniente considerando, para cada intervalo, su marca de clase (el punto medio). Agreguemos a la tabla anterior las marcas de clase de cada intervalo, indicadas como *x'*:

Tabla 12

Tiempo de reacción (en segundos)	x'	f
1,0-1,5	1,25	5
1,5-2,0	1,75	7
2,0-2,5	2,25	6
2,5-3,0	2,75	3
3,0-3,5	3,25	8
3,5-4,0	3,75	5
Total		34

Ahora puede usarse el método anterior para calcular la media, tomando las marcas de clase como los valores de la variable:

$$\bar{x} = \frac{1{,}25 * 5 + 1{,}75 * 7 + 2{,}25 * 6 + 2{,}75 * 3 + 3{,}25 * 8 + 3{,}75 * 5}{34} = 2{,}5$$

Resulta así que el tiempo promedio de reacción es de 2,5 segundos.

Si bien la media es una medida muy valiosa para resumir un conjunto de datos, a veces se hace un uso abusivo de ella, al aplicarla a variables que no tienen el nivel de medición adecuado para autorizar su uso. Un ejemplo de esto es el caso de las calificaciones escolares, que solo permiten ordenar a los alumnos según los resultados, pero que no implican la proporcionalidad de los valores (quien obtiene 10 no sabe el doble que quien obtiene 5). Aun así, es habitual que se calcule incorrectamente el "promedio de las notas".

Medidas no centrales

Los cuartiles

Si la variable tiene un nivel de medición ordinal o superior, entonces podemos usar el mismo razonamiento con el que definimos la mediana para hacer cortes más finos en una distribución de frecuencia. Así, si la mediana nos indica el valor de la variable que deja por debajo la mitad de los casos, es lícito preguntar también por el valor que deja por debajo un cuarto de los casos, o también el que deja por debajo las tres cuartas partes de las observaciones. Estos puntos de corte se denominan respectivamente: primer cuartil y tercer cuartil.

El primer cuartil es el valor de la variable que deja por debajo un cuarto, o el 25% del total de observaciones.

El tercer cuartil es el valor que deja por debajo las tres cuartas partes o el 75% del total de observaciones. Como se ve, tanto el modo de cálculo como la interpretación son análogos a la mediana. Veamos su aplicación a los ejemplos anteriores:

Tabla 10

Número de materias aprobadas	f	F
0	30	30
1	150	180
2	200	380
3	300	680
4	250	930
5	200	1130
6	20	1150
Total	1150	

Para encontrar el primer cuartil será ahora necesario buscar un cuarto del total de casos: 287,5 ((1/4)*1150). La pregunta ahora es ¿cuál es la primera frecuencia acumulada que supera a 287,5? se trata de 380, que corresponde al valor 2 y éste es entonces el primer cuartil. Leemos así que un cuarto del total de alumnos tiene dos materias aprobadas o menos. También puede decirse que el 25% de los alumnos aprobó dos materias o menos. Si se trata de alumnos que han cursado el primer año de la carrera, este grupo, que no llegó a aprobar tres materias, estaría presentando dificultades particulares y sobre él se podría prever una intervención. La utilidad de esta medida es que nos informa que ese grupo con problemas constituye el 25% del total de alumnos que cursó primer año.

El **primer cuartil** es el valor de la variable que deja un cuarto (25%) de los casos por debajo y tres cuartos (75%) por encima. Se indica Q_1.

Idéntico razonamiento seguimos para calcular el tercer cuartil: las tres cuartas partes del total es 862,5 ((3/4)*1150). Buscamos luego la primera frecuencia acumulada que supera a ese valor y hallamos que es 930 y que su categoría correspondiente es 4. Entonces el tercer cuartil es 4 materias aprobadas. La lectura será: las tres cuartas partes (o el 75%) de los alumnos aprobó cuatro materias o menos. Esto último implica que el 25% restante aprobó más de cuatro materias.

El **tercer cuartil** es el valor de la variable que deja tres cuartos (75%) de los casos por debajo y un cuarto (25%) por encima. Se indica Q_3.

Cuando se trata de distribuciones con categorías agrupadas, procedemos como antes con una leve modificación en la fórmula:

Tabla 11

Tiempo de reacción (en segundos)	f	F
1,0-1,5	5	5
1,5-2,0	7	12
2,0-2,5	6	18
2,5-3,0	3	21
3,0-3,5	8	29
3,5-4,0	5	34
Total	34	

Para el primer cuartil debemos hallar la primera frecuencia que supera a un cuarto de las observaciones, de las 34, un cuarto es 8,5 y la primera frecuencia mayor que ese número es 12, por lo que el primer cuartil se encontrará en la categoría 1,5-2,0. Para interpolar en valor exacto usamos una expresión equivalente a la de la mediana:

$$Q_1 = l_i + i * \left(\frac{\frac{n}{4} - f_d}{f_p} \right)$$

En la que solo hemos cambiado $\frac{n}{2}$ por $\frac{n}{4}$ y lo demás mantiene el mismo significado. Aplicándola a estos datos resulta:

$$Q_1 = 1{,}5 + 0{,}5 * \left(\frac{8{,}5 - 5}{7} \right) = 1{,}5 + 0{,}5 * (0{,}5) = 1{,}75$$

Leemos en resultado como: el 25% de los sujetos reaccionó en un tiempo de 1,75s o menos.

Para el tercer cuartil la fórmula se transforma en:

$$Q_3 = l_i + i * \left(\frac{\frac{3 * n}{4} - f_d}{f_p} \right)$$

Cuyo cambio consiste en que hacemos $\frac{3*n}{4}$ en lugar de $\frac{n}{4}$ y manteniendo el resto de los símbolos con el mismo significado.
Usando esta expresión, verifique que para la distribución de los tiempos de reacción, el tercer cuartil es 3,28.

No hemos hecho mención a un “segundo cuartil”, que sería el valor de la variable que acumula las dos cuartas partes de los casos, pero como las dos cuartas partes es la mitad, se trata simplemente de la mediana $Q_2 = M_{dn}$.

Los percentiles

De manera equivalente pueden definirse cortes en otros puntos de la distribución, los más frecuentemente usados, por su generalidad, se conocen como percentiles. Se trata de valores de la variable que dejan por debajo (acumulan) distintos porcentajes de casos.

El **percentil *r*** de una distribución es el valor de la variable que deja el r por ciento de los casos por debajo de él y (1-r) por ciento de los casos por encima. Se indica P_r.

Así por ejemplo, el percentil 10 (indicado como P_{10}) es el valor de la variable que acumula el 10% de las observaciones. Se representa de modo general un percentil dado como P_r en el que r indica el porcentaje del que se trata. La expresión para el cálculo de cualquier percentil es:

$$P_r = l_i + i * \left(\frac{\frac{r}{100} * n - f_d}{f_p} \right)$$

Son fáciles de observar las siguientes equivalencias:

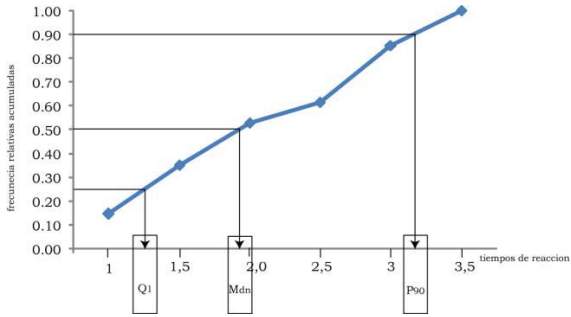
$$Q_1 = P_{25}$$
$$M_{dn} = P_{50}$$
$$Q_3 = P_{75}$$

También suelen mencionarse, en algunas publicaciones, otros puntos de corte, como por ejemplo los quintiles. Esta medida representa valores que acumulan quintos (20%) de la distribución. La equivalencia es la siguiente:

Quintil	Equivale a:
Primero	P_{20}
Segundo	P_{40}
Tercero	P_{60}
Cuarto	P_{80}

Obtención gráfica de los percentiles

Todas las medidas que hacen uso de las frecuencias acumuladas (mediana, cuartiles, percentiles) pueden obtenerse de manera aproximada a través del gráfico de frecuencias acumuladas, llamado ojiva. Veamos la forma de identificar el primer cuartil, la mediana y el percentil 90 en el ejemplo de los tiempos de reacción:



En el gráfico se ve que, si se ubica en el eje de las ordenadas la frecuencia relativa acumulada de 0,25 (un cuarto) y desde allí se llega hasta la ojiva, entonces en el eje horizontal (valores de la variable) se encuentra el primer cuartil. Para hallar la mediana se empieza ubicando en el eje vertical la frecuencia relativa acumulada de 0,50 (la mitad) y luego se la identifica en el eje horizontal. Del mismo modo con el percentil 90, ubicando ahora la frecuencia relativa acumulada de 0,90.

Obtención informática de las medidas de posición

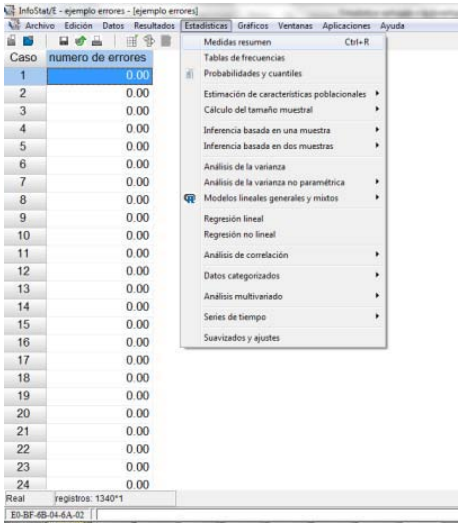
El paquete estadístico InfoStat® obtiene estas medidas de tendencia central directamente desde los datos sin necesariamente mostrar la distribuciones de frecuencia aunque lo usual es también solicitarlas, a fin de mostrar más detalles sobre la variable que se observa. Por ejemplo, para el número de errores cometidos por un conjunto de 1339 personas al responder a una prueba de diez preguntas, se obtiene:

Frecuencias absolutas

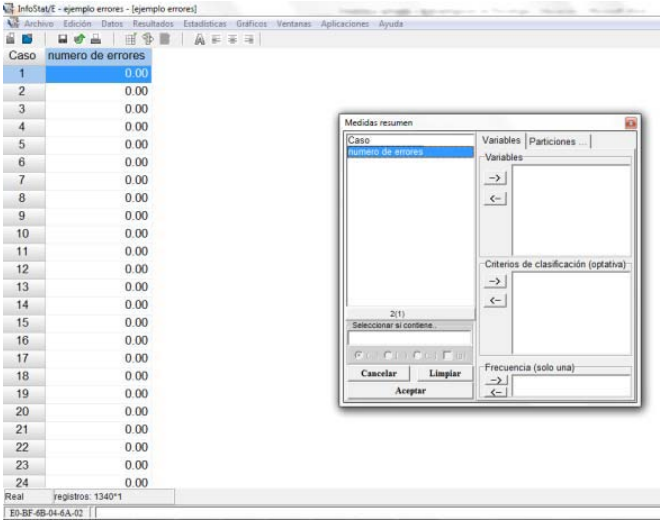
número de errores	Total	Porcentaje
0,00	129	9,63
1,00	185	13,82
2,00	244	18,22
3,00	221	16,50
4,00	220	16,43
5,00	150	11,20
6,00	82	6,12
7,00	58	4,33
8,00	50	3,73
Total	1339	100,00

Las siguientes pantallas muestran la secuencia de comandos para solicitar a InfoStat® la descripción de esta variable:

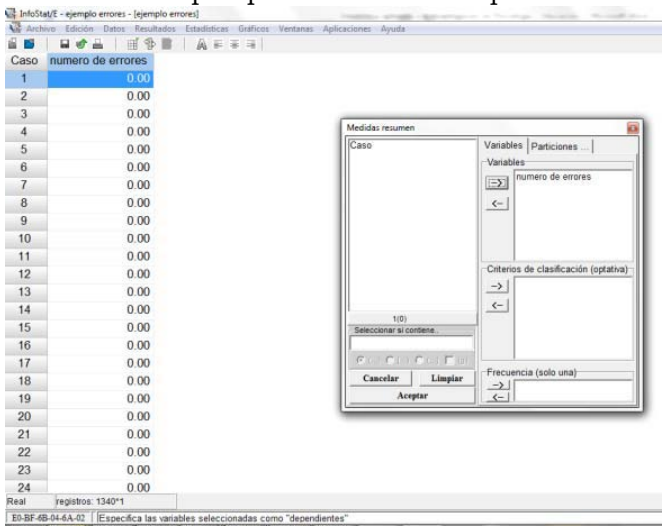
1. Seleccionamos “medidas resumen” en el menú “estadísticas”



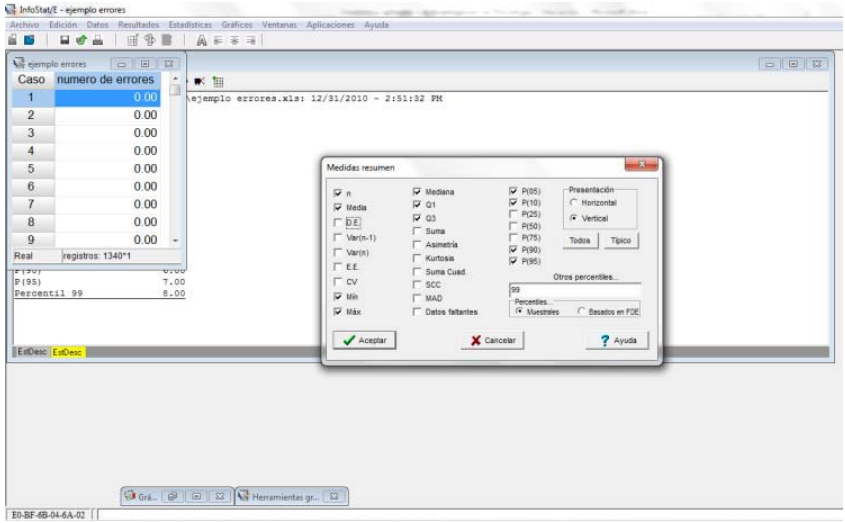
2. Aparece el listado de variables de la base, en este ejemplo solo *número de errores*



3. La ingresamos al recuadro de la derecha para indicar que es de esa variable de la que queremos su descripción



4. Tenemos a nuestra disposición las medidas que elegiremos para solicitar. En este caso elegimos: *n* (número de casos), la media, los valores mínimo y máximo, la mediana, los cuartiles 1 y 3, los percentiles 5, 10, 90, 95 y 99. Adicionalmente, por cuestiones de comodidad, pedimos presentación vertical de los resultados:



Luego de “aceptar” obtenemos la siguiente salida:

Medidas resumen

Resumen	numero de errores
n	1339
Media	3,18
Mín	0,00
Máx	8,00
Mediana	3,00
Q1	2,00
Q3	5,00
P(05)	0,00
P(10)	1,00
P(90)	6,00
P(95)	7,00
Percentil 99	8,00

Esta descripción de los datos puede leerse del siguiente modo:
Cuando 1339 personas respondieron a una prueba de diez preguntas, el número promedio de preguntas respondidas de manera equivocada fue de 3,18. La cantidad de errores cometidos fue desde cero hasta ocho. La mitad de los participantes cometió tres errores o menos, el 25% cometió dos o menos y el 75% hasta cinco. El 5% no cometió ningún error, el 10% cometió uno o ninguno. El 90% cometió seis errores o menos, el 95% cometió siete o menos. También decirse, usando el P_{10}, que solo el 10% cometió más de seis errores, o con el P_{95}, que solo el 5% cometió más de siete.
Verifique que resulta clara toda la lectura de las medidas de posición obtenidas en la salida.

La forma de la distribución

La media es una medida muy completa como resumen de los datos, ya que los considera a todos con la frecuencia de cada uno. Opera como un punto de equilibrio en un conjunto de datos. Sin embargo esto puede ser una dificultad en algunos tipos de distribución. Consideremos el siguiente ejemplo simple:

Tabla 13

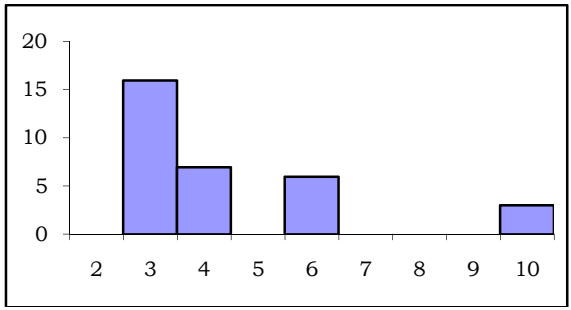
x	f
3	16
4	7
6	6
10	3
total	32

Estos datos muestran una marcada concentración en el valor 3, donde se encuentra la mitad de las observaciones. El resto de los valores son superiores y hay uno extremo, el 10, que tiene poca frecuencia: hay solo tres observaciones con ese valor. Veamos cuál es el efecto de esta forma de distribuirse de los datos. La media es:

$$\bar{x} = \frac{3 * 16 + 4 * 7 + 6 * 6 + 10 * 3}{32} = 4{,}44$$

A pesar de la concentración en 3 que recién mencionamos, la media es superior a 4, que es un resultado contrario a lo que intuitivamente esperaríamos, porque habríamos supuesto que se ubicaría más cerca de 3, ya que 3 parece ser un valor muy "representativo" de esta distribución, sin embargo, la media da un número bastante más grande. Esto se debe a la presencia de valores extremos, en este ejemplo el 10. Aunque este número tiene poca frecuencia, su efecto es de "tirar de la media" hacia valores más grandes. Esto sucede siempre con la media y proviene de su característica de tener en cuenta todos los valores de la distribución. Por esa razón, cuando la distribución se presente como la anterior, la media no será una buena medida de centralidad.
Este inconveniente de la media aparece a menudo en las discusiones salariales por sectores. A menudo se escucha que no se justifica un aumento porque el salario promedio de todos los empleados del sector es de *X pesos*. Argumento al que se contrapone (expresado de diferentes maneras) que ese promedio incluye al personal que tiene salarios muy altos. Se trata de distribuciones asimétricas, que tienen la mayor parte de los casos con salarios bajos o intermedios y unos pocos casos con salarios muy superiores, por eso cuando se calcula la media se obtiene un resultado que representa mal al conjunto de datos.

Exploremos un poco más la forma de la distribución de la tabla 13, su histograma es el siguiente:



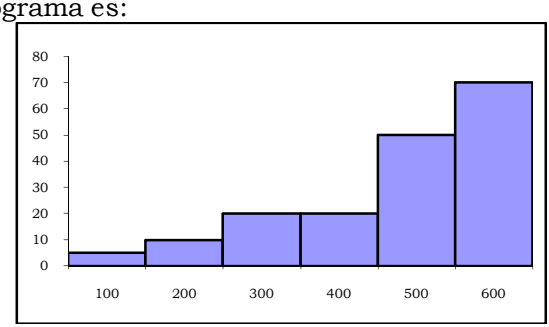
El gráfico muestra el carácter atípico del valor 10, que aparece muy alejado de la parte principal de la distribución. Decimos en este caso que la distribución es asimétrica. La **asimetría** de una distribución se indica señalando hacia dónde se sitúan los valores extremos. Si, como en este ejemplo, el (o los) valor(es) extremo(s) es mayor que la mayor parte de los datos, la asimetría es **hacia la derecha**.

La asimetría puede ser en sentido opuesto, en el caso que haya observaciones particularmente pequeñas y en ese caso tendremos una distribución **asimétrica hacia la izquierda**. Como en el ejemplo siguiente:

Tabla 14

X	f
100	5
200	10
300	20
400	20
500	50
600	70
Total	175

Cuyo histograma es:



Aquí los valores extremos se encuentran por debajo del grupo principal de datos y la media se inclinará hacia valores más pequeños de los más centrales. Así, aunque la mayoría de los casos se encuentra entre 500 y 600, la media es:

$$\bar{x} = \frac{100 * 5 + 200 * 10 + 300 * 20 + 400 * 20 + 500 * 50 + 600 * 70}{175} = 477{,}14$$

Un resultado que está por debajo de esos valores que concentran muchos casos. Decimos ahora que la asimetría es hacia la izquierda.

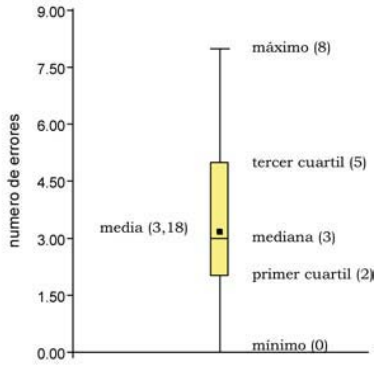
La asimetría puede evaluarse directamente a partir de las medidas de centralidad, ya que la posición relativa de la media y la mediana indican hacia dónde ésta sucede. Cuando la media y la mediana coinciden, la distribución es simétrica, es decir carece de asimetría. Si la media supera a la mediana, se trata de una distribución asimétrica a la derecha y si la media es menor que la mediana, la asimetría será hacia la izquierda.

Posición relativa de la media y la mediana	Asimetría de la distribución
$\bar{x} = M_{dn}$	Simétrica
$\bar{x} > M_{dn}$	Asimétrica a la derecha
$\bar{x} < M_{dn}$	Asimétrica a la izquierda

Una distribución es **simétrica** si la media coincide con la mediana. La distribución se llama asimétrica a la derecha si la media es mayor que la mediana, y asimétrica a la izquierda si la media es menor que la mediana.

Representación gráfica

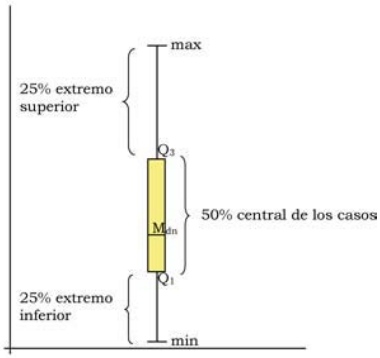
Disponemos de un gráfico que puede resumir de manera muy compacta la información sobre una distribución de frecuencias que hemos generado hasta aquí. Se llama diagrama de caja, o también diagrama de caja y bigotes y se conoce muy frecuentemente como Box-plot, propuesto por John Tukey en 1977. Aplicado al ejemplo de la variable *número de errores*, relevada sobre 1339 casos, que mostramos antes, ofrece.



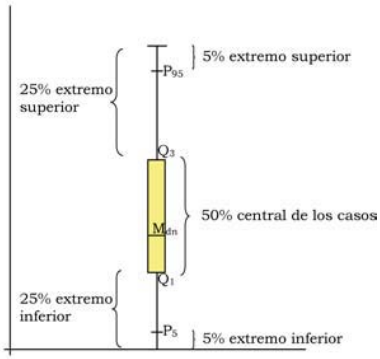
Este gráfico representa sobre el eje vertical los valores de la variable y muestra una "caja" delimitada por los cuartiles 1 y 3. Según la definición de los cuartiles, esa caja contiene al 50% central de los casos. Dentro de la caja se muestra la mediana y la media, lo que nos ofrece una idea gráfica de la asimetría de la distribución.

Además de la caja, se ven dos segmentos que se extienden hasta los valores máximo y mínimo de la distribución. Estos segmentos (llamados a veces "bigotes") muestran el campo completo de variación de la variable y completan la idea acerca de la asimetría de la distribución. En el apartado siguiente veremos que también es posible apreciar visualmente la distancia que hay entre las diferentes observaciones.

De manera general entonces, el box-plot permite apreciar gráficamente la distribución de los casos:



En InfoStat®, también pueden pedirse los percentiles 5 y 95, con lo que la información que ofrece el gráfico es más rica:

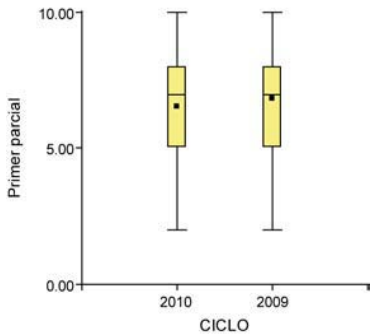


La evaluación exploratoria de los datos puede usarse también de manera comparativa. Por ejemplo, las notas del primer parcial de Psicoestadística en 2009 y 2010 son descriptas en la siguiente salida InfoStat®:

Medidas resumen

CICLO	Resumen	Primer parcial
2009.00	n	1345.00
2009.00	Media	6.82
2009.00	Mín	2.00
2009.00	Máx	10.00
2009.00	Mediana	7.00
2009.00	Q1	5.00
2009.00	Q3	8.00
2010.00	n	1210.00
2010.00	Media	6.52
2010.00	Mín	2.00
2010.00	Máx	10.00
2010.00	Mediana	7.00
2010.00	Q1	5.00
2010.00	Q3	8.00

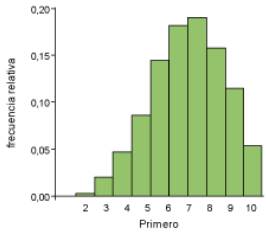
La variable *ciclo* es la que separa los dos grupos que se comparan y corresponde a quienes cursaron en 2009 y 2010.
La representación gráfica de esta descripción, a través de box-plots, es la siguiente:

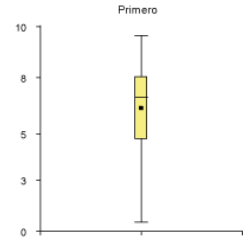


En el que se aprecia la gran similitud en los resultados obtenidos por los alumnos.

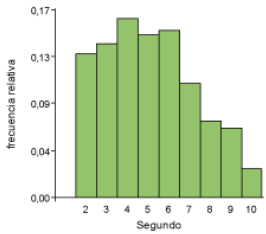
Para comparar las formas de representación de los resultados, a continuación vemos el histograma junto al box-plot de las notas de los dos primeros parciales del año 2010:

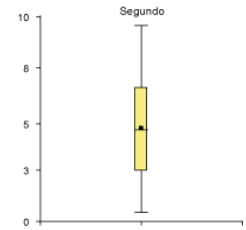
Resumen	Primero
n	1210
Media	6,52
Mín	2
Máx	10
Mediana	7
Q1	5
Q3	8





Resumen	Segundo
n	1231
Media	5,07
Mín	2
Máx	10
Mediana	5
Q1	3
Q3	7





Se puede observar la diferente asimetría de las dos distribuciones expresada en el sesgo de los histogramas y en la distinta longitud de los bigotes en los box-plot.

Medidas de dispersión

Además de indicar alrededor de qué valores se distribuyen los datos, también es necesario indicar si se encuentran concentrados alrededor de esos valores (si son cercanos a ellos) o dispersos (si están alejados). Por ejemplo, un promedio de 20 sesiones de psicoterapia puede provenir de cuatro casos que utilizaron 18, 19, 21 y 22 sesiones o de otros cuatro que hayan insumido 5, 10, 30 y 35 sesiones. En la primer situación las cuatro observaciones son cercanas a la media, están concentradas a su alrededor, mientras que en la segunda están lejos, dispersas. Diremos que en el primer caso la distribución es homogénea o que presenta poca dispersión y en el segundo que es heterogénea o que presenta mucha dispersión.

Conocer esto tiene importancia para poder evaluar la calidad de las medidas de centralidad, en particular de la media. Esto es así porque en una distribución muy dispersa, la media será un promedio de valores muy diferentes entre sí y no será tan fiel a los datos como si estos valores fueran similares. La media de 20 sesiones del primer ejemplo es una mejor medida resumen que la misma media de 20 del segundo, porque la primera representa mejor los datos de origen. Debido a esto, decimos

que en la primera de las situaciones del ejemplo, la media es más *representativa* de los datos de los que proviene.
Nos ocuparemos ahora del modo en que puede medirse esa dispersión, cómo transformarla en una medida resumen que indique brevemente si los datos están dispersos o concentrados.

Recorrido
Una primera aproximación al problema es la de considerar la distancia que hay entre los valores extremos, entre el primero y el último. Si usamos este procedimiento en el ejemplo anterior vemos que en la primera distribución hay 4 unidades entre la primera y la última observación (de 18 a 22) y en la segunda hay 30 unidades de extremo a extremo (de 5 a 35). Por lo que ésta sería una buena medida de la dispersión. Esta medida se llama recorrido, se indica con la letra R y la expresión formal de su cálculo es:

$$R = x_{max} - x_{min}$$

Donde x_{max} y x_{min} representan a los valores máximo y mínimo respectivamente.

Se llama **recorrido** de una distribución a la diferencia entre los valores máximo y mínimo de la variable. Se indica *R*.

Cuando la distribución tiene más casos, el recorrido es insuficiente como medida de dispersión, ya que está determinado solo por los valores extremos. Por ejemplo, las dos siguientes series tienen la misma media, igual a 8:

2, 8, 8, 8, 8, 8, 14
7, 8, 8, 8, 8, 8, 9

El recorrido vale 12 para la primera (*R=14–2*) y 2 para la segunda (R=9–7) es una diferencia muy acentuada aunque las dos distribuciones solo difieren en los valores extremos. Dicho de otra manera, si sucede que hay un caso (o unos pocos) que tiene un valor excepcionalmente alto (o bajo), el recorrido dará un valor alto, indicando gran dispersión, lo que nos puede hacer pensar que todos los datos están dispersos. Por esa razón se dice que es una medida "gruesa" de la variabilidad de los datos.

Amplitud intercuartílica
Un modo de afinar la calidad de esta medida es la de tomar la distancia que hay no ya entre los valores extremos sino entre los cuartiles primero y tercero. La medida que usa esta distancia se llama amplitud intercuartílica y es simplemente la diferencia entre el tercer cuartil y el primero:

$$AIQ = Q_3 - Q_1$$

Si bien tampoco es ésta una medida que considere todas las observaciones —ya que solo tiene en cuenta los dos cuartiles—, es mejor que el recorrido, porque deja de lado los valores extremos, aquellos que pertenecen al 25% más bajo y al 25% más alto de la distribución.

La **amplitud intercuartílica** es la diferencia entre los cuartiles tercero y primero. Se indica *AIQ*.

Medidas de dispersión basadas en la media

Las medidas de variabilidad que más se usan son las que tienen en cuenta todas las observaciones, es decir aquellas que están basadas en la media. Una manera de ver si el conjunto de datos está concentrado o disperso, consiste en observar la distancia de la media a la que se encuentra cada observación, luego esas distancias individuales pueden promediarse y tener una idea global de qué tan lejos están los casos del promedio. Intentemos hacer eso y veamos qué limitación aparece.

Tomemos un conjunto pequeño de datos, presentado en serie simple:

5, 7, 9, 11

La media es 8, como lo es la mediana. Aunque no hay modo, ya que todos los valores tienen frecuencia igual a uno, la distribución es simétrica. Hemos elegido así el ejemplo solo para darle simplicidad, no es una condición necesaria para lo que sigue.

Tomemos ahora las distancias a las que cada observación se encuentra de la media, restando a cada una de ellas el valor 8 (la media):

x_i	5	7	9	11
$x_i - \bar{x}$	-3	-1	1	3

Si sumamos todas las diferencias $x_i - \bar{x}$, el resultado es cero (-3-1+1+3=0); además, éstas son simétricas, como efecto de la forma de la distribución original. Pero el hecho que la suma sea cero no depende de la forma de la distribución, sino que es una propiedad de la media. Por ser la media un punto de equilibrio

entre las observaciones, las que se distancian por encima de ella están compensadas por las que lo hacen por debajo[22].

Los valores $x_i - \bar{x}$ se llaman desvíos, que indican cuánto se aleja cada observación de la media. Como vemos pueden ser positivos o negativos según se trate de observaciones que superen a la media o que estén por debajo de ella. Acabamos de ver también que su suma vale cero, es decir que $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ y que esta es una cualidad de la media, que no depende de los datos[23]. Tan importante es esta propiedad que la usaremos para dar una definición más completa de la media:

La **media** es el valor de la variable que anula la suma de los desvíos en torno suyo.

En el tema que nos ocupa en este momento, el de medición de la variabilidad del conjunto de casos, la consecuencia de esta propiedad es que no será posible usar la suma de los desvíos como indicador de dispersión, ya que da siempre cero, con datos homogéneos o heterogéneos. A fin de resolver este problema vamos a eliminar el signo, usando el hecho que todo número elevado a una potencia par es positivo, sin importar el signo que haya tenido el número. Elevaremos entonces al cuadrado cada una de los desvíos y así se perderá su signo y ya no será cero la suma de todos ellos.

Usando ese recurso, definimos la varianza[24], a la que simbolizaremos como $V(x)$ o más frecuentemente como s^2 de la siguiente forma:

---

[22] Para ver esto, comparemos con el caso de la serie: 3, 4, 6, 7, 23, 45 con media 14,7. Las diferencias entre cada observación y la media son las siguientes:

x	3	4	6	7	23	45
$x_i - \bar{x}$	-11,7	-10,7	-8,7	-7,7	8,3	30,3

En este caso las diferencias no son simétricas, pero es igualmente cierto que su suma es igual a cero, es decir que están compensadas las diferencias por encima y por debajo de la media.

[23] Puede verse que es así haciendo: $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$, como $\bar{x}$ es una constante, el segundo término es $n * \bar{x}$, igual que el primero, según la definición operativa de la media. Por lo tanto la diferencia es cero.

[24] En este punto aparece la primera diferencia entre cálculos hechos sobre datos de una muestra o de una población. Si estuviésemos trabajando sobre toda la población, la varianza (a la que indicaríamos con otra letra) tendría denominador *n*, en lugar de *n-1*. No podemos explicar la razón de esto aun, habrá que esperar al capítulo de estimación.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Se llama **varianza** de una distribución a la suma de los cuadrados de los desvíos alrededor de la media, dividida por el total de observaciones menos uno. Se indica s^2.

Es una medida muy valiosa de la dispersión que tiene un conjunto de datos, cuanto mayor es, tanto más dispersos éstos se encuentran, es decir, son más heterogéneos. No puede ser negativa, porque es una suma de cuadrados y solo es cero si todos los desvíos son cero, es decir si todas las observaciones coinciden con la media[25].

Hay tres propiedades de la varianza que señalaremos para su uso posterior:
-La varianza de una constante es cero. Esto resulta claro ya que la varianza mide la dispersión y si todas las observaciones son iguales no hay dispersión:

$$V(k) = 0$$

-La varianza de una constante que multiplica a una variable es la constante elevada al cuadrado multiplicada por la varianza de la variable:

$$V(k * x) = k^2 * V(x)$$

-La varianza de la suma de dos variables independientes es la suma de las varianzas de cada una de ellas:

$$V(x + y) = V(x) + V(y)$$

A los fines de la interpretación, la varianza presenta dos inconvenientes. Uno es que sus unidades están elevadas al cuadrado; por lo que, si medimos *número de errores*, la varianza quedará expresada en *número de errores al cuadrado* una entidad que no tiene significado, como tampoco lo tienen *hijos al cuadrado* o *segundos al cuadrado*, para los tiempos de reacción.
El otro inconveniente es que no tiene límite superior, puede ser muy grande y no tenemos con qué compararla para saber si indica una gran variabilidad o si es grande porque los valores de la variable lo son.
Para resolver el primer inconveniente, definiremos una medida derivada de la varianza, que se denomina desviación estándar (en algunos textos y programas de análisis de datos es llamada

[25] En este caso no hay variabilidad y, en consecuencia, no hay variable, porque el valor asumido es siempre el mismo. Se trata de una constante.

desviación típica). Esta medida, indicada con la letra s se calcula como la raíz cuadrada de la varianza:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

O más simplemente:

$$s = \sqrt{s^2}$$

La **desviación estándar** es la raíz cuadrada de la varianza. Se indica *s*.

Ahora, por el sencillo trámite de introducir una raíz cuadrada, las unidades de s son las mismas que las de la variable original y no hay problemas con la interpretación del valor.

Para hacer frente al problema de la magnitud de la varianza —que sigue siéndolo para la desviación estándar— definimos una medida relativa de la dispersión, el coeficiente de variación, indicado como CV y que no es sino el cociente entre la desviación estándar y la media:

$$CV = \frac{s}{\bar{x}} * 100$$

Esta medida carece de unidades, porque la media tiene las mismas que las de la desviación estándar, por lo que se trata de una medida relativa de la dispersión. Indica la importancia relativa de la desviación estándar respecto de la media. El factor 100 que acompaña al cociente cumple la función de expresarlo como porcentaje, por comodidad para la lectura.

El **coeficiente de variación** expresa de manera relativa la dispersión, midiendo el peso de la desviación estándar comparado con la media. Se indica *CV*.

Conocer la dispersión de una distribución de frecuencias es muy necesario para poder decidir si la media es una medida adecuada para resumir los datos, y esto no sucede si hay mucha dispersión. Para aclarar esto veamos un ejemplo: sea un grupo de seis alumnos que hacen una prueba y que obtienen las siguientes notas: 2, 2, 2, 2, 10, 10. Si calculamos la media obtenemos 4,7. Este número no representa lo que sucede con los seis alumnos, quienes tuvieron resultados muy dispares: cuatro de ellos obtuvieron 2 y los otros dos, 10. Si calculamos el CV, resultado es 100%, un valor muy elevado, indicativo que la

media no es una medida adecuada para sintetizar al conjunto de datos. Muchas de las críticas mal fundadas hacia la Estadística se equivocan porque "muestran" el uso de la media cuando no corresponde calcularla.

En la práctica se considera que si el coeficiente de variación es menor al 10%, la distribución tiene poca dispersión y entonces podemos confiar en la media como medida de centralidad y tratarla como representativa de los datos que resume.

Calcularemos por única vez las medidas de dispersión de manera manual para un pequeño conjunto de datos, a fin de seguir las operaciones que involucra. Se trata de seis pacientes diagnosticados de depresión a partir de cinco o más de los síntomas que indica el manual DSM IV[26] y que para cada uno de ellos observamos (como variable) el número de síntomas que llevaron al diagnóstico:

[26] Presencia de cinco (o más) de los siguientes síntomas durante un período de 2 semanas, que representan un cambio respecto a la actividad previa; uno de los síntomas debe ser:
1. Estado de ánimo depresivo la mayor parte del día, casi cada día según lo indica el propio sujeto (p. ej., se siente triste o vacío) o la observación realizada por otros (p. ej., llanto). En los niños y adolescentes el estado de ánimo puede ser irritable
2. Disminución acusada del interés o de la capacidad para el placer en todas o casi todas las actividades, la mayor parte del día, casi cada día (según refiere el propio sujeto u observan los demás)
3. Pérdida importante de peso sin hacer régimen o aumento de peso (p. ej., un cambio de más del 5 % del peso corporal en 1 mes), o pérdida o aumento del apetito casi cada día. Nota: En niños hay que valorar el fracaso en lograr los aumentos de peso esperables
4. Insomnio o hipersomnia (sueño excesivo) casi cada día.
5. Agitación o enlentecimiento psicomotores casi cada día (observable por los demás, no meras sensaciones de inquietud o de estar enlentecido)
6. Fatiga o pérdida de energía casi cada día
7. Sentimientos de inutilidad o de culpa excesivos o inapropiados (que pueden ser delirantes) casi cada día (no los simples autorreproches o culpabilidad por el hecho de estar enfermo)
8. Disminución de la capacidad para pensar o concentrarse, o indecisión, casi cada día (ya sea una atribución subjetiva o una observación ajena)
9. Pensamientos recurrentes de muerte (no sólo temor a la muerte), ideación suicida recurrente sin un plan específico o una tentativa de suicidio o un plan específico para suicidarse

Paciente	x_i (número de síntomas)	$x_i - \bar{x}$ (desvíos)	$(x_i - \bar{x})^2$ (cuadrados de los desvíos)
1	5	-2	4
2	6	-1	1
3	6	-1	1
4	8	1	1
5	8	1	1
6	9	2	4

$$\bar{x} = \frac{5+6+6+8+8+9}{6} = 7$$

$$\sum_{i=1}^{6}(x_i - \bar{x})^2 = 4+1+1+1+1+4 = 12$$

$$s^2 = \frac{\sum_{i=1}^{6}(x_i - \bar{x})^2}{n-1} = \frac{12}{6-1} = 2{,}4$$

$$s = \sqrt{s^2} = \sqrt{2{,}4} = 1{,}55$$

$$CV = \frac{s}{\bar{x}} * 100 = \frac{1{,}55}{7} * 100 = 22{,}13\%$$

La lectura de este resultado es que para el conjunto de seis personas a las que se observa, el número promedio de síntomas a través de los cuales es diagnosticada la depresión es de siete. Sin embargo este número de síntomas es bastante variable según los pacientes y, seguramente también según los terapeutas.

Obtención informática de medidas de dispersión
Si la serie de datos del ejemplo anterior es cargada en InfoStat®, las medidas descriptivas se solicitan en el menú: Estadísticas, Medidas resumen. Luego de seleccionar la variable que se describirá, se eligen las medidas, el formato de la salida es el siguiente:

Estadística descriptiva

Variable	n	Media	D.E.	Var(n-1)	CV
NUMERO DE SINTOMAS	6	7,00	1,55	2,40	22,13

Que también puede pedirse presentado de manera vertical:

Estadística descriptiva

Resumen	NUMERO DE SINTOMAS
n	6,00
Media	7,00
D.E.	1,55
Var(n-1)	2,40
CV	22,13

En la salida, *n* es la cantidad de casos, *D.E.* se refiere a la desviación estándar, *Var(n-1)* es la varianza, en la que la indicación *(n-1)* señala que se ha calculado con denominador *(n-1)*, por lo que se trata de la varianza muestral. *CV* el coeficiente de variación, expresado como porcentaje. Esta salida puede leerse "Sobre un total de seis pacientes diagnosticados como depresivos, el número promedio de síntomas presentes en que se basó el diagnóstico fue de 7. Las observaciones son levemente heterogéneas, ya que el coeficiente de variación es superior al 20%. De aquí puede concluirse que el número de síntomas que apoyan el diagnóstico de depresión es bastante variable según los pacientes."

Box-plots y dispersión

La observación del diagrama de caja (box-plot) nos da también indicios acerca de la dispersión de la variable que se analiza. Cuando la caja es larga estaremos en presencia de distribuciones muy dispersas en la parte central, mientras que si la caja es corta, será indicador de una concentración de datos en la parte central de la distribución. La longitud de las colas (o bigotes) nos dirá la mayor o menor concentración de los datos en las zonas extremas. Como dijimos antes, el box-plot es un gráfico que ayuda a explorar los datos, a hacerse una idea inicial de la distribución y esto puede ser muy valioso cuando se trata de interpretarlos, porque permite sugerir hipótesis que expliquen la distribución que se observa.

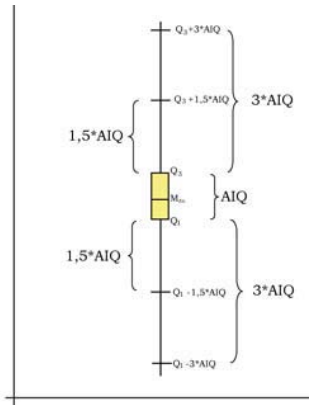
Haciendo uso de la amplitud intercuartílica estableceremos criterios para detectar valores que destaquen por alejarse sustancialmente del grupo mayoritario. Se trata de mediciones atípicas o excepcionalmente extremas, ya porque sean excesivamente grandes o pequeñas. La identificación de estos valores es importante en la etapa exploratoria de los datos porque obliga a determinar qué sucede con esos casos. Puede tratarse de un error de medición o bien de un sujeto (o unos

pocos) que se aparta de manera excepcional del grupo y que merece un análisis más detallado y particularizado.

Tukey (1970) sugiere tratar como "lejanas" a las observaciones que se encuentren a más de una amplitud intercuartílica y media (1,5*AIQ) por debajo del primer cuartil o por encima del tercero, pero a menos de tres veces la amplitud intercuartílica (3*AIQ). Además, aquellas observaciones que estén más allá de tres AIQ por debajo del primer cuartil o por encima del tercero se denominan "muy lejanas". Este criterio determina entonces zonas en las que pueden hallarse las observaciones y según en cuál de ellas se encuentren, se las identifica como "cercanas", "lejanas" o "muy lejanas". Las zonas son las siguientes:

1. Cercanas: Entre Q_1 y Q_1-1,5*AIQ o entre Q_3 y Q_3+1,5*AIQ
2. Lejanas: Entre Q_1-1,5*AIQ y Q_1-3*AIQ o entre Q_3+1,5*AIQ y Q_3+3*AIQ
3. Muy lejanas: Menores que Q_1-3*AIQ o mayores que Q_3+3*AIQ

La división en zonas puede verse más claramente en un box-plot:



En este esquema hemos tomado la distancia entre los cuartiles tercero y primero (la amplitud intercuartílica) como unidad de medida, y consideramos una vez y media esa medida (1,5*AIQ) y tres veces esa medida (3*AIQ) como puntos de corte para decidir cuándo una observación se aleja excepcionalmente del grupo.

Medida de la dispersión cuando no hay distancias

Todo lo indicado hasta el momento acerca de la variabilidad ha necesitado de la medición de la distancia entre las observaciones: desde el comienzo hablamos de cercanía o lejanía entre los datos. Por lo tanto estas medidas, desde el recorrido hasta el coeficiente de variación, solo tienen sentido si la variable es de nivel intervalar

o proporcional. Si la variable tiene nivel nominal u ordinal habremos de medir su variabilidad de un modo diferente. En estos casos cambia un poco el significado de la variabilidad, ya que estaremos en presencia de una variable más dispersa cuanto más *equitativamente* se distribuya el total de observaciones entre las distintas categorías. Por ejemplo, si 100 individuos son clasificados según cómo sea su rendimiento en: muy bueno, bueno, regular, insatisfactorio; la distribución tendrá más dispersión si 25 de ellos se encuentran en cada categoría que si la gran mayoría está en una sola. La distribución:

Tabla 15

Rendimiento	f	f'
Muy bueno	25	0,25
Bueno	25	0,25
Regular	25	0,25
Insatisfactorio	25	0,25
Total	100	1,00

Tiene más dispersión que esta otra:

Tabla 16

Rendimiento	f	f'
Muy bueno	5	0,05
Bueno	80	0,80
Regular	5	0,05
Insatisfactorio	10	0,10
Total	100	1,00

¿Por qué? Porque en la segunda, los casos están concentrados en una categoría, mientras que en la primera se dispersan entre todas. Notemos que ahora tendremos más dispersión cuanto más parecidas sean las frecuencias entre sí. Esto puede parecer contradictorio con lo indicado para variables cuantitativas, pero allí la mayor dispersión viene dada por la mayor disparidad entre los valores de las variables, que no puede evaluarse con variables nominales u ordinales.

Esta forma de considerar la dispersión equivale a la idea de incertidumbre. Supongamos que conocemos que la distribución del rendimiento es como lo muestra la tabla 15 y que debemos "adivinar" cuál es el rendimiento de una persona elegida al azar. No tenemos ninguna razón para creer de manera preferencial que la persona sea de rendimiento muy bueno, bueno, regular o insatisfactorio; ya que todos son igualmente posibles. En esta situación, la incertidumbre es completa. Por el contrario, si supiéramos que la distribución es la que muestra la tabla 16,

tenderíamos con justa razón a creer que la persona elegida al azar tiene rendimiento bueno, ya que es bastante más probable que pertenezca a esa categoría que a otra. Diremos que aquí tenemos menos incertidumbre.
La medida para expresar de manera sintética esta dispersión es:

$$H(x) = -\sum_{i=1}^{k} f_i^{'} * log f_i^{'}$$

El cálculo consiste en multiplicar cada frecuencia relativa por su propio logaritmo y sumar para todas las categorías. El resultado de la sumatoria siempre es negativo, por lo que la fórmula incluye un signo menos para volverlo positivo. Este coeficiente expresa en un solo número la magnitud de la dispersión. Cuanto más pequeña sea esta medida, tanto menos dispersa (o más concentrada) será la distribución de la variable que se analiza.

Aplicado a las dos tablas de más arriba obtenemos

para la tabla 15:

$$H(x) = -(0{,}25 * log 0{,}25 + 0{,}25 * log 0{,}25 + 0{,}25 * log 0{,}25 + 0{,}25 * log 0{,}25) = -(-0{,}60) = 0{,}60$$

y, para la tabla 16:

$$H(x) = -(0{,}05 * log 0{,}05 + 0{,}80 * log 0{,}80 + 0{,}05 * log 0{,}05 + 0{,}10 * log 0{,}10) = -(-0{,}31) = 0{,}31$$

Así, a la distribución en la que las frecuencias están más concentradas, es decir la que tiene menor dispersión (tabla 16), le corresponde un menor valor de *H(x)*.

El individuo en relación a su grupo

Nos interesa plantear aquí un uso muy frecuente en evaluación psicológica y educativa de las medidas que acabamos de ver y que permite decidir si un valor particular está cerca o lejos del promedio, o bien si se sitúa o no en los extremos de una distribución. Así formulado el problema puede parecer muy elemental, porque puede "verse" si un número está cerca o lejos de otro. Si sabemos que una persona tiene dos metros de estatura, no necesitamos hacer cuentas para saber que es alto, más alto que la mayoría de las personas. Sin embargo, en el caso de medidas menos familiares, y como veremos en los ejemplos siguientes, a veces resulta difícil hacer juicios de distancia sobre valores absolutos.

Si sabemos que en una prueba de memoria con un puntaje máximo de 100 puntos, una persona logró 80 puntos, ¿estamos autorizados para decir que obtuvo un puntaje alto? La respuesta es no, porque no sabemos qué puntajes obtuvieron las demás personas que hicieron la prueba. Si la media del grupo completo hubiese sido 60 puntos, entonces 80 sería un valor elevado, pero si la media hubiese sido de 85, entonces el caso que estamos considerando se encontraría por debajo del promedio. Más aun, si el promedio fuese 60 y la mayoría de los evaluados hubiese obtenido puntajes cercanos a 60 (poca variabilidad), entonces el valor 80 podría considerarse como muy elevado. Solo conocer su puntaje individual no nos dice nada acerca de la posición de un sujeto particular.
Otro ejemplo: nos informan que un niño obtuvo un puntaje bruto de 85 en la escala de desarrollo infantil de Bayley, no tenemos, en principio ningún criterio para decidir si ese puntaje es alto o bajo.

Para situaciones como éstas, muy frecuentes en evaluaciones psicológicas y educativas, será necesario conocer cuál es la posición *relativa* que un puntaje ocupa respecto del conjunto completo de observaciones.
Supongamos que se aplica una prueba de ortografía a una muestra de alumnos de tercer grado y que el promedio de errores es 10 ($\bar{x} = 10$ errores) y que la desviación estándar es de 4 ($s = 4$ errores). Si un alumno comete 6 errores ($x = 6$ errores), podemos decir que cometió menos errores que el promedio del grupo. El cálculo de la diferencia entre x y $\bar{x}$ da -4 errores ($x - \bar{x} = 6 - 10 = -4$), este resultado nos informa que este alumno se ubica a 4 errores por debajo del promedio (por debajo queda expresado en el signo menos el resultado). Ésta es una medida concreta, ya que expresa el número de errores que separan al alumno del comportamiento resumido del grupo (expresado en la media); dicho de otra manera, estamos considerando los valores absolutos. Si ahora a esta diferencia la dividimos por la desviación estándar obtenemos -1 (procedente de $\frac{-4}{4}$), que ya no tiene unidades, es un número abstracto. Como la desviación estándar es de 4 puntos y el alumno se encuentra a cuatro puntos de la media, esto equivale a decir que el alumno se encuentra "a una desviación estándar por debajo del promedio". La operación que hemos hecho ha sido la de restar al valor particular (de ese alumno) la media y dividir el resultado en la desviación estándar, hemos calculado lo siguiente:

$$\frac{x - \bar{x}}{s}$$

Este número, que como dijimos no tiene unidades, es diferente para cada valor de x y mide la distancia a la que se encuentra una observación particular (x) de la media ($\bar{x}$), expresada como fracción de la desviación estándar (s). Decimos que se trata de una medida *estandarizada* del alejamiento que tiene una observación particular del promedio del conjunto de observaciones.
Hemos así expresado la posición del alumno respecto del grupo al que pertenece de manera relativa, en términos de desviaciones estándar.
La variable que resulta de esta operación se llama **desvío estándar**, ya que se trata de un desvío (calculado en la diferencia $x - \bar{x}$) expresado como cantidad de desviaciones estándar. Se utiliza la letra z para indicarla, así:

$$z = \frac{x - \bar{x}}{s}$$

Debido a que la letra z se utiliza de manera universal para indicar este valor, es también conocido como **puntaje z** o **puntuación z**. Esta nueva variable tiene media igual a cero y desviación estándar igual a uno[27].
Volvamos sobre el ejemplo del número de síntomas en que se basa el diagnóstico de depresión, cuya media fue de 7 y su desviación estándar de 1,55.

Paciente	x_i (número de síntomas)	$x_i - \bar{x}$ (desvíos)	z (desvíos estándar)
1	5	-2	-1,29
2	6	-1	-0,65
3	6	-1	-0,65
4	8	1	0,65
5	8	1	0,65
6	9	2	1,29

La última columna proviene de haber dividido cada desvío en la desviación estándar (1,55). Los desvíos indican a cuántas unidades de la variable (en este caso *número de síntomas*) se ubica cada caso del promedio. Los desvíos estándar indican a cuántas desviaciones estándar se encuentra cada caso del

[27] Dado que $\bar{x}$ y s son constantes, aplicando las propiedades de la media, resulta:
$\bar{z} = \overline{\left(\frac{x-\bar{x}}{s}\right)} = \frac{\bar{x}-\bar{x}}{s} = 0.$
También haciendo uso de las propiedades de la varianza, la de z es:
$V(z) = V\left(\frac{x-\bar{x}}{s}\right) = \frac{1}{s^2}\left(V(x) - V(\bar{x})\right) = \frac{1}{s^2}(V(x) - 0) = \frac{1}{s^2}V(x) = 1,$
ya que la media es constante por lo que tiene varianza nula, y $V(x) = s^2$

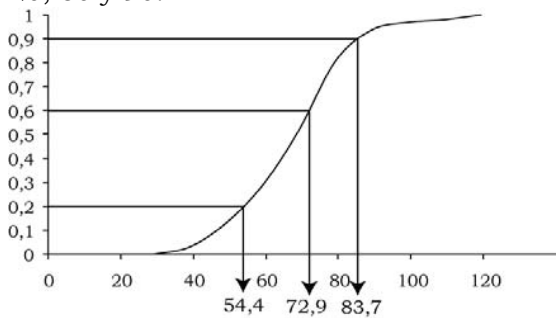
promedio. El primer paciente está a 1,29 desviaciones estándar por debajo del promedio, el tercero está a 0,65 desviaciones estándar por debajo del promedio, etc.

Cuando se trata de variables de nivel ordinal también es posible ubicar de manera relativa cada valor de la variable, aunque no puedan medirse distancias. Es así porque en esas variables podemos calcular percentiles e indicar a qué percentil corresponde cada valor. Antes vimos el modo de señalar gráficamente la ubicación de los percentiles, allí buscamos de identificar el valor de la variable que corresponde, por ejemplo, al percentil 90 ó a cualquier otro. Podemos hacer ahora el recorrido inverso: dado un valor de la variable ¿a qué percentil corresponde?
Consideremos los siguientes puntajes brutos obtenidos en una prueba psicológica administrada a una muestra de 310 personas:

Tabla 17

x	f	F	F´
20-29	0	0	0,00
30-39	10	10	0,03
40-49	30	40	0,13
50-59	50	90	0,29
60-69	70	160	0,52
70-79	90	250	0,81
80-89	40	290	0,94
90-99	10	300	0,97
100-109	5	305	0,98
110-119	5	310	1,00
Total	310		

Los percentiles de esta distribución pueden obtenerse gráficamente usando la ojiva. Por ejemplo, hallamos los percentiles 20, 60 y 90:



Procediendo del mismo modo, hacemos la siguiente correspondencia de puntajes brutos a percentiles:

Percentil	X
10	47,0
20	54,4
30	60,4
40	64,9
50	69,3
60	72,9
70	76,3
80	79,8
90	87,3

La tabla nos informa sobre los valores de la variable donde se divide cada 10% del total de casos. Usando la definición de los percentiles diremos que:
- El 10% de los sujetos obtuvo 47 puntos o menos
- El 20% obtuvo 54,4 puntos o menos y así para el resto.

Con esta información sabemos que si una persona obtuvo 50 puntos, tiene un puntaje muy bajo, porque supera a menos del 20% del grupo. O dicho de otra manera, más del 80% de las personas alcanzaron puntajes más altos que él. Por el contrario si alguien obtuvo 88 puntos, tiene un puntaje muy alto, ya que supera al percentil 90, con lo que menos del 10% del grupo lo supera. O bien, él supera a más del 90%.

De este modo, la construcción de una tabla en la que se indica el valor de la variable (el puntaje en la prueba) correspondiente a cada percentil, permite conocer si un puntaje dado se ubica en algún extremo de la distribución (si es excepcionalmente elevado o bajo) o si es un valor intermedio.

Estas tablas de correspondencia entre valores absolutos (o puntajes brutos) y los correspondientes valores relativos pueden también construirse usando los desvíos estándar, transformando cada valor observado en su puntuación *z*. Para el ejemplo de la tabla 17 necesitamos calcular la media y la desviación estándar, que dan: $\bar{x} = 68{,}4$ y $s = 15{,}7$. Con esto podemos indicar los puntajes z que corresponden a cada puntaje bruto, con la transformación $z = \frac{x-\bar{x}}{s}$.

Para la marca de clase del primer puntaje bruto (25), el puntaje *z* que le corresponde es:

$$z = \frac{25-68,4}{15,7} = -2,8.$$

Repetimos esta operación para cada puntaje bruto y obtenemos la tabla de correspondencias:

Intervalo de puntajes brutos	Puntaje z
20-29	-2,8
30-39	-2,1
40-49	-1,5
50-59	-0,9
60-69	-0,2
70-79	0,4
80-89	1,1
90-99	1,7
100-109	2,3
110-119	3,0

Una tabla de ese tipo (ya sea construida a partir de los percentiles o bien de los puntajes *z*) se conoce como *baremo* y es absolutamente necesario para cualquier tipo de evaluación psicológica o educativa ya que posibilita decidir en qué lugar se encuentra un sujeto dado, respecto de su grupo de referencia, y esto se requiere porque, por ejemplo, para una prueba de inteligencia, un puntaje que es normal para la edad de 13 años, no lo es para los 16. El baremo provee la transformación de puntajes absolutos en puntajes relativos.

Un **baremo** es una tabla de valores transformados que permiten ubicar a un sujeto en relación a su grupo de referencia.

Aplicación a un estudio local[28]

Dentro de la materia Técnicas Psicométricas (segundo año de la Licenciatura en Psicología, UNC), los alumnos han realizado una toma de la escala general del Test de Raven[29] a una muestra de alumnos secundarios y universitarios de la ciudad de Córdoba. Con esos datos, un grupo de alumnos coordinado por un ayudante de la cátedra, se interesó en actualizar el baremo del test, puesto que no existían en ese momento baremos locales y actualizados. Se tomó muestra de 551 alumnos de nivel

[28] Agradecemos a Marcelo Vaiman el aporte de este ejemplo en el que él participó.

[29] O test de matrices progresivas, es una prueba desarrollada por John Raven en 1936 y dirigida a medir las dos componentes principales de la inteligencia general, que Spearman había definido años antes.

secundario y universitario, con edades entre 16 y 17 años, de ambos sexos.
A continuación, los protocolos fueron clasificados por rangos, para cada uno de los cuales se calcularon los percentiles y los puntajes *z* teniendo en cuenta el puntaje total y el de cada serie y se construyeron los baremos. A continuación se presentan en dos versiones, como rangos percentilados y como puntajes *z*:

Puntaje Total (bruto)	Percentil
28	5
39	10
43	25
46	40
48	50
49	60
50	70
51	75
52	80
54	90
55	95
58	99

Intervalo de puntaje bruto	Puntaje *z*
0	-5,83
1 – 2	-5,64
3 – 4	-5,38
5 – 6	-5,13
7 – 8	-4,88
9 – 10	-4,63
11 – 12	-4,37
13 – 14	-4,12
15 – 16	-3,87
17 – 18	-3,61
19 – 20	-3,31
21 – 22	-3,10
23 – 24	-2,86
25 – 26	-2,61
27 – 28	-2,35
29 – 30	-2,10
31 – 32	-1,85
33 – 34	-1,59
35 – 36	-1,35
37 – 38	-1,09
39 – 40	-0,84
41 – 42	-0,59
43 – 44	-0,36
45 – 46	-0,06
47 – 48	0,17
49 – 50	0,42
51 – 52	0,68
53 – 54	0,93
55 – 56	1,18
57 – 58	1,43
59 – 60	1,68

Resumen de las medidas descriptivas definidas en el capítulo

			Nivel de medición mínimo requerido
Medidas de posición	Centrales	Modo	Nominal
		Mediana	Ordinal
		Media	Intervalar
	No centrales	Cuartiles	Ordinal
		Quintiles	
		Percentiles	

			Nivel de medición mínimo requerido
Medidas de dispersión	Entre extremos	Recorrido	Intervalar
	Basada en el orden	Amplitud intercuartílica	Intervalar
	Basadas en la media	Varianza	Intervalar
		Desviación estándar	
		Coeficiente de variación	
	De incertidumbre	Coeficiente de incertidumbre	Nominal

Actividad práctica de repaso 3

1. El Laboratorio de Psicología de nuestra facultad, especializado en el estudio de alcohol, aprendizaje y adolescencia, desarrolló un estudio descriptivo tendiente a evaluar cuales eran las bebidas alcohólicas más consumidas por los adolescentes. Para ello se administró un cuestionario a 384 adolescentes de la ciudad de Córdoba. Los resultados de esta investigación dirigida por Juan Carlos Godoy fueron:

¿Cuál es la bebida que más tomas?

Bebida alcohólica		
No toma		0,323
Cerveza	93	0,242
Vino	29	0,075
Gancia	39	0,102
Fernet	52	0,135
Ron		0,003
Vodka	9	0,023
Sidra o Vino espumoso	35	
Otra bebida	2	0,005
Total	384	1,000

a. Calcule las frecuencias y proporciones faltantes.
b. ¿Cuál es la bebida más consumida por los adolescentes?
c. ¿Cuál es la moda de la distribución?

2. En un estudio sobre Psicología Política dirigido por Silvina Brussino (2007), se administró una escala de Conocimiento Político a una muestra de 299 jóvenes cordobeses. La escala consistía en una serie de preguntas sobre conocimiento cívico y político a la cual los encuestados debían responder. Si la respuesta era correcta se asignaba un valor de 3, si era parcialmente correcta un valor de 2 y si era incorrecta un valor de 1. Algunos de los resultados obtenidos fueron:

Preguntas	Media	Mediana
¿Cuáles son los 3 poderes del Estado?	2,12	3
¿Cuál es el organismo encargado de decidir sobre la constitucionalidad de las leyes?	1,55	1
¿Quiénes son los responsables del nombramiento de Jueces de la Suprema Corte?	1,42	1
¿Cuál es el organismo encargado de promulgar leyes provinciales?	1,57	1
¿Un ciudadano puede asistir a las sesiones del Poder Legislativo?	2,20	3

a. Indique un nombre para las variables que surgen de estas preguntas
b. ¿Cuál es el nivel de medición de las variables?
c. ¿Qué medida descriptiva es más adecuada?
d. ¿Cuáles fueron las preguntas peor respondidas? ¿Cuáles las mejor respondidas?

3. Para la siguiente descripción:

Variable	M_o	M_{dn}	Q_1	P_5	P_{95}	$\bar{x}$
Situación conyugal	2	3	1	1	4	1,7
Cantidad de aplazos	4	4	3	1	8	5,5
Año de la carrera que cursa	1	3	1	1	5	2,8

a. Elija las medidas que puedan interpretarse de acuerdo al nivel de medición de las variables.
b. Redacte una lectura de cada una de las que sea posible.
c. Indique, si corresponde, la simetría o asimetría de las distribuciones.

4. Un grupo de profesores del ingreso universitario realizaron un estudio observando los comportamientos no verbales que los estudiantes realizaban al dar una exposición oral (Medrano y Flores Kanter, 2009). El estudio consistió en puntuar utilizando una escala del 1 al 10 la adecuación de una serie de comportamientos no verbales (manejo de la mirada o uso del espacio, por ejemplo). Las puntuaciones podían variar desde 1 (uso muy inadecuado) hasta 10 (uso muy adecuado). Los resultados obtenidos fueron los siguientes:

Comportamientos No Verbales	Mediana	Rango
Manejo de la mirada	7	4
Uso de gestos	7	5
Postura corporal	7	5
Uso de la sonrisa	7	7
Manejo del espacio	6	7
Velocidad del discurso	7	5
Volumen de la voz	8	4
Manejo del tiempo	8	5

a. ¿Qué significa que el recorrido en el uso de la sonrisa sea mayor que el recorrido observado en el volumen de la voz?
b. ¿Qué otros comportamientos mostraron gran variabilidad? ¿A que podría deberse dicha variabilidad?

5. La siguiente salida InfoStat® describe los resultados de los tres parciales de Psicoestadística en 2010:

Medidas resumen

Resumen	Primero	Segundo	Tercero
n	1312,00	1229,00	1148,00
Media	6,56	5,07	6,28
D.E.	1,92	2,25	2,21
CV	29,31	44,31	35,16
Mín	1,00	1,00	1,00
Máx	10,00	10,00	10,00
Mediana	7,00	5,00	6,00
Q1	5,00	3,00	5,00
Q3	8,00	7,00	8,00
P(05)	3,00	2,00	3,00
P(95)	10,00	9,00	10,00

a. Ofrezca una lectura tan completa como sea posible.
b. Compare los resultados de los tres parciales en su posición y dispersión

Capítulo 4: Relaciones entre variables

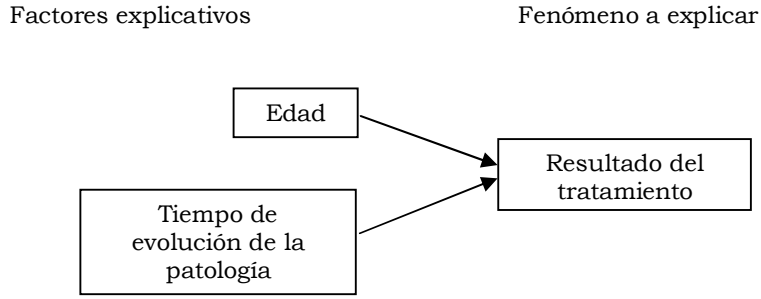
Eduardo Bologna

Hemos trabajado hasta este punto en la descripción de variables observadas o medidas a través de nuestros instrumentos. Continuamos en este capítulo usando datos de provienen de una muestra o de toda la población a la que queremos caracterizar, pero ahora lo haremos con un objetivo que se acerca más a los de las investigaciones en Psicología, Educación y otras Ciencias Sociales. Es así porque este capítulo y el siguiente buscan identificar *relaciones* entre variables: no ya describir cada variable por separado sino reunirlas en relaciones de dos como mínimo, pero que puede incluir a una gran cantidad. Buscar relaciones entre variables es comenzar a transitar el camino de la explicación de los fenómenos que observamos.

Si nos preguntamos, por ejemplo: ¿por qué un tratamiento es exitoso con algunos pacientes diagnosticados de depresión y con otros no? Formularemos hipótesis sobre la respuesta: quizás la edad influya, puede suceder que con pacientes más jóvenes se obtenga mejor resultado que con los de más edad. Razonando así, introducimos otra variable, la *edad*, que aportaría a explicar la razón de los diferentes resultados del tratamiento. La hipótesis está formulada como una relación entre dos variables: se trata de indagar por el efecto que la *edad* (primera variable) tendría sobre el *resultado del tratamiento* (segunda variable). La edad podría ser un **factor explicativo** del resultado del tratamiento.

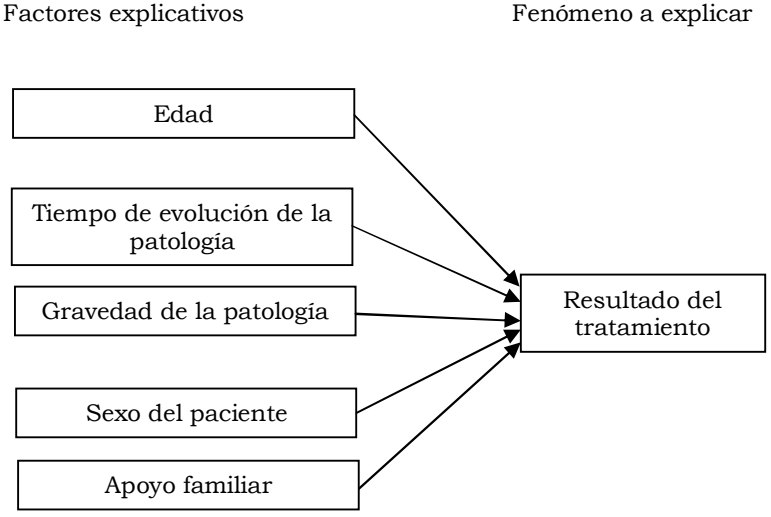
Dentro del mismo ejemplo, también podemos sospechar que quienes han sido diagnosticados más precozmente pueden aprovechar el tratamiento mejor que quienes traen una dolencia de larga data. Aquí la variable que viene a explicar el resultado es el *tiempo de evolución de la enfermedad*. Ahora el *tiempo de evolución de la enfermedad* podría ser otro factor explicativo del *resultado del tratamiento*. Notemos el acento en "podría ser": estas relaciones son hipotéticas, nuestro objetivo será analizar la evidencia que haya a su favor o en su contra. Esquemáticamente la relación se plantea de la siguiente manera:

Esquema 1: Relación hipotética entre dos factores explicativos y el resultado de un tratamiento para la depresión



Estas dos variables son parte de nuestra hipótesis para explicar las diferencias en los resultados que ofrece un determinado tratamiento sobre pacientes diagnosticados de depresión. Puede haber más variables: la gravedad de la depresión, el sexo del paciente (quizás el resultado no sea igual en mujeres que en hombres), el apoyo familiar que el paciente reciba, etc. Tendríamos entonces un esquema explicativo más completo:

Esquema 2: Relación hipotética entre cinco factores explicativos y el resultado de un tratamiento para la depresión



Jamás agotaremos el conjunto de todos los factores explicativos de un fenómeno, porque en última instancia cada caso es único. Los fenómenos que observamos son multicausados, por lo que no puede decirse que una variable *X* sea la causa de otra variable *Y*[30]. Pero lo que sí podemos hacer, y tiene la mayor importancia en investigación, es analizar la importancia relativa de los diferentes factores explicativos.
Ilustremos esto con otro ejemplo: el resultado escolar que alcanzan los alumnos. No hay dudas que cada niño tiene una trayectoria única, que depende de su historia, de su contexto familiar, etc. Supongamos que analizamos el resultado escolar obtenido en primer grado y observamos que algunos cursos tienen docentes tradicionales, que usan los mismos métodos estandarizados de enseñanza desde hace muchos años. Otros cursos tienen docentes que invitan a los alumnos participar, que innovan en los métodos de enseñanza. Luego comparamos el rendimiento de los alumnos en los dos cursos y vemos que los alumnos del primer grupo aprenden más lentamente que los del segundo y que además, los primeros dicen que se aburren yendo a la escuela y los otros no. Esto no sucede con todos los niños: habrá en el primer grupo algunos que aprenden más rápido y que se divierten, así como algunos del segundo grupo tardarán más en aprender. Pero, en general, en promedio, podríamos hallar mejores resultados entre los alumnos que tienen docentes innovadores. Esto nos lleva a indicar que hay evidencia para creer que los docentes innovadores obtienen con sus alumnos mejores resultados que los docentes tradicionales. Pero esto no es para todos los alumnos, sino para la mayoría de ellos. De eso se trata la búsqueda de factores explicativos: en este ejemplo diremos que, de los múltiples factores que explican por qué a algunos chicos les va bien en la escuela y a otros les va mal, el tipo de docente es parte de la explicación.

Las hipótesis son respuestas tentativas a la pregunta formulada como problema de la investigación. Como tales, consisten en el planteamiento de una relación entre, al menos, dos variables. Recordemos que las hipótesis constituyen afirmaciones que se derivan del modelo de análisis que el investigador ha propuesto para explicar una situación dada. Las hipótesis son consecuencias deductivas de la teoría, cuya verificación no es suficiente para validar la teoría, aunque sí para "aportar evidencia en su favor". Además, las hipótesis como tales, rara vez pueden ponerse a prueba de manera directa, son sus

[30] Salvo en el caso de diseños experimentales que permitan tener control sobre el conjunto de variables que participan en el resultado que se observa

consecuencias observables las que permiten la verificación empírica. En cualquier modelo explicativo hipotético participa un número de variables mayor a dos, sin embargo de las hipótesis pueden deducirse relaciones más simples, inicialmente solo de dos variables. En la primera parte de este capítulo nos ocuparemos de relaciones que involucran solo a dos variables y luego avanzaremos hacia la inclusión de otras variables en la contrastación del modelo.

Establecer de manera hipotética una relación entre dos variables equivale a afirmar que, por alguna razón, los cambios de una de ellas van acompañados de cambios en la otra. Pero esto puede suceder de maneras muy diferentes, por ejemplo, el trueno sucede al relámpago, los síntomas de tuberculosis coinciden con la detección del bacilo de Koch, los movimientos sociales se incrementan en tiempos de deterioro económico, las personas abusadas en la infancia son más propensas a la depresión. En algunos de estos ejemplos puede identificarse una secuencia cronológica, señalando cuál de los dos eventos sucede primero, en otros esta distinción no es segura, a veces una variable es la que incide sobre la otra, otras veces es solo una contribución, por último, hay casos en que su ocurrencia conjunta o sucesiva se debe a otras razones. Evitaremos, por ahora hablar de relaciones de causalidad, llegaremos a este concepto hacia el final del capítulo y veremos que debe tratarse con suma cautela.
Con el objetivo de ordenar la gran variedad de formas que pueden asumir las relaciones entre variables, estableceremos algunos criterios de clasificación que, sin ser exhaustivos, nos ayudarán a verlas desde diferentes ópticas.
El modo más usado para observar relaciones entre dos variables consiste en presentar el comportamiento conjunto de ellas a través de tablas o gráficos. Las primeras son más adecuadas para variables con pocas categorías (usualmente nominales), mientras que los gráficos son más pertinentes para mostrar relaciones entre variables métricas. Veamos un ejemplo para ilustrar el primer modo de representación.

Creemos que los niños que han crecido en diferentes tipos de hogar (solo con su madre, solo con su padre, con ambos o con otros parientes) tienen diferentes formas de relacionarse con sus compañeros (con relaciones de liderazgo, sumisión o rebeldía). En el lenguaje de las relaciones entre variables, estaríamos proponiendo que existe **asociación** entre el tipo de hogar en que el niño crece (con las cuatro categorías mencionadas) y el tipo de relación que mantiene con sus pares. Presentamos conjuntamente esas dos variables de este modo:

Esquema 3: Disposición de las variables y sus categorías para analizar la relación entre dos variables

	Relación con los pares		
Tipo de hogar	Sumisión	Rebeldía	Liderazgo
Monoparental materno			
Monoparental paterno			
Nuclear			
Extendido			

Cuando se distribuyen los datos en las celdas, se obtiene una tabla bivariada (porque contiene dos variables), que también se llama tabla de contingencia o tabla de distribución conjunta. Las celdas del interior de la tabla llevarán, cuando los datos sean recolectados, la cantidad de niños que se encuentren en cada coincidencia de categorías. Si nuestra hipótesis afirmara que los niños provenientes de hogares nucleares son más propensos a ser líderes, esperaríamos una concentración relativa de casos en la celda correspondiente a "hogar nuclear" – "liderazgo", hipótesis que luego deberemos confrontar con la información recogida.

Una **tabla bivariada** o **tabla de contingencia** o **tabla de distribución conjunta** es un arreglo con tantas filas (horizontales) como categorías tenga una de las variables y tantas columnas (verticales) como categorías tenga la otra variable.

A este arreglo se agrega una fila y una columna adicionales que corresponden a los totales de cada categoría. A los fines de usar un lenguaje común, en la tabla llamaremos filas a la líneas horizontales y columnas a las verticales e identificaremos la **dimensión** de la tabla indicando cuántas filas tiene y cuantas columnas, en este orden. En el ejemplo anterior, la dimensión de la tabla es cuatro por tres, porque tiene cuatro filas y tres columnas correspondientes a la cantidad de categorías de cada una de las dos variables.

La **dimensión** de la tabla se indica como *f X c*, donde *f* es el número de categorías de la variable que está en las filas y *c* es el número de categorías de la variable que está en las columnas.

La celda en la que, bajo la hipótesis indicada, esperaríamos una mayor concentración *relativa* de casos corresponde entonces a la

tercera fila y tercera columna. Con *f* indicaremos la frecuencia y con el subíndice la celda a que corresponde, así, f_{ij} será la cantidad de casos en la celda que corresponde a la fila *i* y a la columna *j* simultáneamente. La frecuencia de la celda de "nuclear-liderazgo" será indicada entonces como f_{33}. Para aclarar la presentación de la tabla, se agregan una fila y una columna en la que se indica el total de casos de cada una de ellas, que se llaman **marginales** de fila y de columna. La notación será:
$f_{i.}$ ("efe i-punto") para los marginales de fila
$f_{.j}$ ("efe punto-jota") para los de columnas
$f_{..}$ ("efe punto-punto") para el total general.

Con esa notación, $f_{3.}$ indicará el total de niños que crecieron en hogares nucleares y $f_{.2}$ el total de quienes se vinculan con rebeldía con sus compañeros. La tabla anterior resulta:

Esquema 4: Disposición de las frecuencias para el análisis de la relación entre dos variables.

	Relación con los pares			
Tipo de hogar	Sumisión	Rebeldía	Liderazgo	Total
Monoparental materno	f_{11}	f_{12}	f_{13}	$f_{1.}$
Monoparental paterno	f_{21}	f_{22}	f_{23}	$f_{2.}$
Nuclear	f_{31}	f_{32}	f_{33}	$f_{3.}$
Extendido	f_{41}	f_{42}	f_{43}	$f_{4.}$
Total	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{..}$

Si hemos recogido datos sobre estas características de los niños, la tabla podría quedar así:

Tabla 1: Alumnos de escuelas primarias por relación con sus pares según tipo de hogar.

	Relación con los pares			
Tipo de hogar	Sumisión	Rebeldía	Liderazgo	Total
Monoparental materno	20	30	50	100
Monoparental paterno	10	40	15	65
Nuclear	5	10	25	40
Extendido	30	20	10	60
Total	65	100	100	265

Esta tabla dice que se han observado un total de 265 niños y se ha registrado el tipo de hogar en que crecieron y la forma en que se relacionan con sus compañeros. De los 265:
100 provienen de hogares monoparentales maternos,
65 de monoparentales paternos,
40 de nucleares y
60 de hogares extendidos.
Los mismos niños se relacionan con sus compañeros:
65 de ellos con sumisión,
100 con rebeldía y
100 con liderazgo.
Todas estas son lecturas de las frecuencias marginales. Marginales de fila las del tipo de hogar y marginales de columna las de la forma de la relación.

Se llama **frecuencias marginales de fila** a las frecuencias absolutas de las categorías de la variable que se ubica en las filas.

Las **frecuencias marginales de columna** son las frecuencias absolutas de las categorías de la variable ubicada en las columnas.

Las frecuencias de las celdas, que se llaman **frecuencias conjuntas** se leen: 20 de los niños observados crecieron en hogares monoparentales maternos y se relacionan con sus compañeros con sumisión y del mismo modo el resto de las frecuencias conjuntas. Ellas indican la cantidad de casos que reúnen al mismo tiempo las dos condiciones que se indican en la fila y en la columna.

Las **frecuencias conjuntas** indican la cantidad de casos que corresponden simultáneamente a una determinada categoría de la variable de las filas y una categoría de la variable de columnas.

Del mismo modo en que trabajamos con las tablas de distribución de frecuencia de una sola variable (capítulo 2), podemos transformar todas estas frecuencias absolutas en relativas, por el simple procedimiento de dividirlas en el total general. Resulta así:

Tabla 2: Alumnos de escuelas primarias por relación con sus pares según tipo de hogar, frecuencias relativas

	Relación con los pares			
Tipo de hogar	Sumisión	Rebeldía	Liderazgo	Total
Monoparental materno	0,08	0,11	0,19	**0,38**
Monoparental paterno	0,04	**0,15**	0,06	0,25
Nuclear	0,02	0,04	0,09	0,15
Extendido	0,11	0,08	0,04	0,23
Total	**0,25**	0,38	0,38	1,00

Leemos así las frecuencias que están destacadas en la tabla:
-El 15% del total de alumnos observados proviene de hogares monoparentales paternos y se relacionan con sus pares con rebeldía. Es f'_{22}
-Un 25% del total se relaciona con sumisión, sin considerar el tipo de hogar del que provengan. Es $f'_{1.}$
-Un 38% proviene de hogares monoparentales maternos, sin tener en cuenta de qué manera se relacionan con sus pares. Es $f'_{.1}$
La primera de estas frecuencias relativas es conjunta, las otras dos marginales. Verifique que queda bien clara la notación usada en cada caso y que pueden leerse las demás frecuencias relativas de la tabla.

Una clasificación en referencia al tiempo
Como señalamos al principio hay relaciones en las que resulta posible identificar a una de las variables como previa a la presencia de la otra, o a un evento como anterior a la ocurrencia del otro. Así, el trueno siempre sucede luego del relámpago, si tenemos la oportunidad de oírlo. Aún cuando no podamos establecer la causa de la relación entre los dos eventos, no tenemos dudas en señalar a uno como anterior al otro. Los malos tratos sufridos durante la niñez son anteriores (en la historia del sujeto) a la eventual manifestación adulta de conductas antisociales. De modo que si nos interrogáramos sobre la existencia de una relación entre estos dos eventos, ubicaríamos a los malos tratos como variable anterior, aunque solo fuera porque su manifestación es temporalmente previa. Así sucede también al buscar una relación entre los resultados de un examen de ingreso a la universidad y el rendimiento posterior de los alumnos.
Es importante indicar a esta altura que no estamos suponiendo que la relación exista, nos encontramos en el momento del planteo de las

hipótesis; bien puede suceder que, luego del análisis de los datos, encontremos que la relación no es válida, que no se sostiene, en fin, que las observaciones no avalan una asociación entre malos tratos infantiles y conducta antisocial, o que el resultado de examen de ingreso no se relaciona con el desempeño posterior, pero esto no invalida que, en la relación que proponíamos, una variable sea tratada como anterior a la otra.

Así como en ciertos casos es posible anticipar el orden (sea lógico o cronológico) en que se presentan las variables que constituyen una relación, hay algunas situaciones en que esto es muy difícil, o imposible y otras en las que no tiene ningún interés. Una relación que ilustra el primer caso es la relación entre el comportamiento infantil y el trato que recibe de sus padres. Puede interpretarse a los niños revoltosos como respondiendo a la escasa atención que le brindan sus padres, o leer la forma en que los padres tratan a sus hijos como consecuencia de la mala conducta de estos últimos. En este ejemplo se ve que el orden en que se establezcan las variables que se busca relacionar está influido por la posición teórica que el investigador asuma.

Otros casos en los que no tiene interés mencionar qué variable es anterior y cual posterior son típicos de los estudios descriptivos, en los que interesa mostrar cómo se distribuyen ciertas variables, y no qué relación puede haber entre ellas. Así, una distribución de la población por sexo y edad como la de la tabla siguiente:

Tabla 3: Departamento Capital, Provincia de Córdoba. Población por sexo según grupos de edad. Año 2001

Grupos de edad	Sexo		Total
	Varones	Mujeres	
0-4	56.913	55.053	111.966
5-9	57.471	56.073	113.544
10-14	55.564	54.394	109.958
15-19	55.581	55.834	111.415
20-24	67.519	69.727	137.246
25-29	53.736	54.667	108.403
30-34	42.209	43.852	86.061
35-39	36.910	39.894	76.804
40-44	34.681	38.243	72.924
45-49	31.879	36.634	68.513
50-54	30.780	36.187	66.967
55-59	24.485	29.448	53.933
60-64	19.914	25.038	44.952
65-69	16.485	22.387	38.872
70-74	13.858	20.831	34.689
75-79	8.816	15.318	24.134
80-84	4.423	9.471	13.894
85-89	1.827	5.355	7.182
90-94	600	1.906	2.506
95-99	120	448	568
100 y más	8	43	51
Total	**613.779**	**670.803**	**1.284.582**

Fuente: INDEC (2009)

Solo pretende describir a la población y no tiene sentido preguntar qué variable es prioritaria a la otra o cuál depende de cuál.

Las relaciones en que no es posible o no interesa señalar qué variable es anterior, se llaman **simétricas** o de variación conjunta o de covariación, con ellas simplemente se indica que las variables están correlacionadas. Queriendo decir en este caso que lo que se observa es que ambas varían simultáneamente sin determinar cuál es la que podría preceder a la otra.

Otro ejemplo de este tipo de relaciones es la que puede plantearse entre las calificaciones que los alumnos obtienen en dos materias que cursan simultáneamente; si encontramos que a aquellos alumnos que les va bien en Epistemología también obtienen buenas notas en Biología, no creeremos que un resultado incida en el otro, solamente *describiremos* que varían conjuntamente. Si luego nos interesamos por avanzar en un estudio explicativo iremos a buscar otras variables que den cuenta de esta covariación.

Una relación entre dos variables es **simétrica** cuando es de variación conjunta y no puede identificarse a una variable como previa a la otra

Por el contrario, aquellas relaciones en las que puede identificarse a una variable como anterior a otra se denominan **asimétricas**, es decir, no es lo mismo planearlas en un sentido que en otro. Una de las variables (la anterior) se llama antecedente y la otra (posterior) consecuente. En algunos contextos (sobre todo en el diseño experimental) estas variables se denominan independiente y dependiente respectivamente.
Puede observarse que una variable cambia a continuación de la otra (en sentido temporal) pero esto no nos autoriza a decir que cambia a causa de la otra, como resulta claro en el ejemplo del relámpago y el trueno. Que la relación sea asimétrica no implica que una variable sea ni la causa, ni un factor explicativo, de la otra.
A la inversa, en los estudios explicativos la relación debe ser asimétrica, porque se busca identificar factores que anticipen determinados eventos. Por eso decimos que la asimetría es una condición necesaria de la causalidad, pero no suficiente.

Una relación entre dos variables es **asimétrica** cuando una de las variables antecede (lógica o cronológicamente) a la otra y puede identificarse a una como antecedente y a la otra como consecuente.

La dirección de la relación
Cuando las variables que se ponen en juego en una relación tienen un nivel de medición superior al nominal, resulta posible hacer juicios de orden entre sus categorías, con lo que es posible indicar si los valores van creciendo o decreciendo. Ya sea que se trate de una relación simétrica o asimétrica, si las variables tienen nivel ordinal o superior, resulta de interés plantear la **dirección** de la relación. Se trata de otro criterio para clasificar relaciones entre variables: si a cambios ascendentes (crecientes) de una variable se siguen cambios

ascendentes de la otra, llamamos a la relación **directa**. Si, por el contrario, un crecimiento de una de las variables va acompañado de una disminución en los valores de la otra, la denominaremos **inversa**. Cuando se espera que la relación entre dos variables sea directa o inversa para toda la serie de categorías, decimos que la relación es monótona.
Por ejemplo, puede plantearse, de manera hipotética, la relación entre los años de educación y el salario. Las personas que han asistido más años a instituciones educativas tienden, en promedio, a tener ingresos más altos que quienes asistieron menos tiempo. La hipótesis anticipa una relación directa entre la escolarización y los ingresos.

Una relación entre dos variables medidas a nivel ordinal o superior es **directa** si cuando los valores de una de ella aumentan, también aumentan los de la otra.

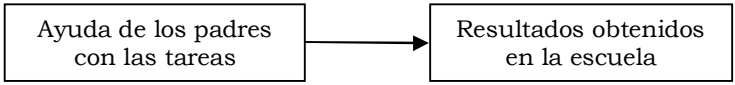
Análogamente:

Se llama **inversa** a la relación entre dos variables de nivel ordinal o superior en la que los incrementos en los valores de una de ellas van acompañados de disminuciones en los valores de la otra

Se explicita en estas definiciones que la clasificación solo tiene sentido si puede hablarse de aumento o disminución, es decir, si es factible realizar juicios de orden entre las categorías de las variables. Por eso es que este criterio requiere, para su aplicación, que ambas variables tengan por lo menos nivel ordinal.
Por ejemplo, la *calificación* que se obtiene en un examen (variable consecuente, de nivel ordinal) puede tener relación directa con las *horas dedicadas* a estudiarla (variable antecedente, de nivel proporcional). Lo que equivale a decir que quienes estudian más horas tenderían a obtener calificaciones más altas.
Si en otro ejemplo, se formula como hipótesis que el *tipo de escuela* secundaria (variable antecedente, de nivel nominal) a la que los alumnos asistieron tiene relación con el *rendimiento* que alcanzan en su carrera universitaria (variable consecuente, de nivel ordinal), no es posible establecer la dirección de esta relación, porque no se cumple que ambas variables sean al menos ordinales.

Un ejemplo
La expresión “si los padres los ayudan con las tareas a los chicos les va mucho mejor en la escuela” equivale a decir que la ayuda que los

padres les dan a sus hijos está relacionada con mucha intensidad con el rendimiento en la escuela. La primera variable, la ayuda es la antecedente, que puede ser de nivel nominal, con categorías: *ayuda*, *no ayuda*; o bien ordinal, con categorías: *ayuda siempre*, *casi siempre*, *pocas veces*, *nunca*. La segunda variable, el rendimiento, es la consecuente, y sus categorías podrían ser: *rendimiento alto*, *medio*, *bajo*. El esquema de la relación será:



Y la tabla que reúna los datos para verificar esta relación podrá tener dimensión 2X3, con forma:

Esquema 5: Disposición de las variables para analizar la relación entre la ayuda que los padres dan a sus hijos y el rendimiento que alcanzan en la escuela

	Rendimiento			Total
Ayuda	Alto	Medio	Bajo	
Si				
No				
Total				

O bien, considerando a la ayuda como ordinal, en una tabla de 4X3

Esquema 6: Disposición de las variables para analizar la relación entre la ayuda que los padres dan a sus hijos y el rendimiento que alcanzan en la escuela

	Rendimiento			Total
Ayuda	Alto	Medio	Bajo	
Siempre				
Casi siempre				
Pocas veces				
Nunca				
Total				

Planteada de este modo, se trata de una relación asimétrica, ya que suponemos que es la ayuda (antecedente) la que incide sobre el resultado (consecuente). Si vemos el esquema 6, puede considerarse la dirección (en el anterior no ¿por qué?) y la formularíamos como directa, es decir que, cuanto mayor sea ayuda que los padres aportan, tanto mejores serán los resultados. Esto está dentro de la hipótesis, aún no hemos recogido datos para avalarla o refutarla. Llegar a

conocer si la ayuda de los padres contribuye en gran medida o escasamente a los resultados en la escuela, es un problema de la intensidad de la relación, que solo podrá responderse a posteriori, una vez que los datos están recolectados.

La intensidad

Sea que se trate de relaciones simétricas o asimétricas y que pueda o no decidirse sobre la dirección, siempre es posible (y tiene mucho interés hacerlo) evaluar la **intensidad** en que se manifiesta la relación entre las variables a partir de los datos de nuestras observaciones. Esta medida de la relación se corresponde con la idea intuitiva de "X tiene mucha influencia en Y", la idea de mucha o poca influencia, es la de intensidad de la relación. Cuando hay muchos factores explicativos para un fenómeno —como en el Esquema 2—, es muy importante poder saber qué factores inciden en mayor o menor medida en el fenómeno y a eso se responde indicando la intensidad de cada relación.

La intensidad o grado de la relación puede también aplicarse a relaciones simétricas. En ese caso, la intensidad mide qué tan a menudo los cambios de una de las variables se ven acompañados de cambios en la otra. Como sucede con las relaciones simétricas, se trata de un resultado descriptivo, no explicativo.

La **intensidad de una relación**[31] es una medida de la fuerza con que los cambios en una variable afectan los cambios en la otra (si es una relación asimétrica) o bien, de la frecuencia con que los cambios de una variable acompañan a los de la otra (si se trata de una relación simétrica).

La evaluación de esta intensidad puede lograrse, en una primera aproximación, observando la distribución conjunta de las dos variables. En la medida que cierta combinación de categorías de una y otra variable concentren la mayor parte de los casos, estaremos en presencia de relaciones más fuertes o de mayor intensidad. Los siguientes son resultados de un estudio que relaciona el tipo de docente con el rendimiento de sus alumnos:

[31] No es posible ofrecer una definición más precisa ya que, como veremos a lo largo de la materia, según el modo en que se mida la intensidad, es decir, según el coeficiente que se use, es diferente el aspecto de la relación que se tiene en cuenta.

Tabla 4: Alumnos primarios por rendimiento según tipo de docente, frecuencias absolutas

tipo de docente	rendimiento			Total
	alto	medio	bajo	
Autoritario	5	35	50	90
Democrático	260	40	10	310
Total	265	75	60	400

InfoStat® ofrece la tabla de contingencia así:

Tablas de contingencia

Frecuencias absolutas
En columnas: rendimiento

tipo docente	1,00	2,00	3,00	Total
1,00	5	35	50	90
2,00	260	40	10	310
Total	265	75	60	400

En la están codificados los tipos de docente como 1 y 2 y del mismo modo los rendimientos de los alumnos.

Al pedir las frecuencias relativas al total obtenemos:

Tabla 5: Alumnos primarios por rendimiento según tipo de docente, frecuencias relativas al total

tipo de docente	rendimiento			Total
	alto	medio	bajo	
Autoritario	0,01	0,09	0,13	0,22
Democrático	0,65	0,10	0,03	0,78
Total	0,66	0,19	0,15	1,00

Leemos las frecuencias conjuntas y las marginales como en la tabla 2. Por ejemplo:
El 10% de los alumnos tuvo docente democrático y rendimiento medio.
El 15% de los alumnos tuvo rendimiento bajo (sin importar qué tipo de docente tuvo).
El 22% tuvo docente autoritario (sin importar cuál fue su rendimiento).

Observando la distribución de los casos de la tabla 4 (frecuencia absolutas) parece haber una tendencia a que los docentes autoritarios tengan alumnos con menor rendimiento. Supondremos que los alumnos han sido distribuidos al azar entre los docentes de los dos tipos, es decir que no sucedió que a los docentes autoritarios "les tocaron" malos alumnos. Por lo tanto nuestra pregunta es acerca del efecto que tendría el tipo de docente (antecedente) sobre el rendimiento de los alumnos (consecuente).
De los 90 alumnos que tuvieron docente autoritario, la mayor parte (50) muestra bajo rendimiento; por el contrario, de los 310 alumnos que tuvieron docente "democrático", la mayoría (260) tiene rendimiento alto, por lo que podría afirmarse que no es igual el rendimiento que se observa en alumnos con docentes de un tipo o de otro.
A fin de eliminar los efectos de las cantidades diferentes de alumnos con docentes autoritarios y democráticos, la relación puede verse mejor si se calculan porcentajes en cada celda. Tomando como totales las cantidades de alumnos que tienen docentes autoritarios y que los tienen democráticos, se obtiene:

Tabla 6: Alumnos primarios por rendimiento según tipo de docente, frecuencias relativas por filas

Tipo de docente	Rendimiento			Total
	Alto	Medio	Bajo	
Autoritario	0,06	0,39	0,55	1,00
Democrático	0,84	0,13	0,03	1,00
Total	0,66	0,19	0,15	1,00

Sobre esta tabla puede afirmarse que, del total de alumnos que tuvieron docentes autoritarios, un 55% ha tenido bajo rendimiento. Por el contrario solo un 3% de los alumnos con docente democrático acusa rendimiento bajo. En el otro extremo, los autoritarios tienen un 6% de alumnos de rendimiento alto frente a un 84% de los democráticos.

Otra manera de leer los porcentajes es comparando con los marginales: sin considerar el tipo de docente, el 66% de los alumnos tuvo rendimiento alto. Cuando se considera al grupo de alumnos que tuvo docentes democráticos, este porcentaje sube al 84%, mientras que, en el grupo de quienes tuvieron docente autoritario, solo el 6% tuvo alto rendimiento.

La ventaja de usar porcentajes (o frecuencias relativas) es que las comparaciones se hacen sobre los mismos totales, es "como si" hubiese 100 alumnos con docente autoritario y 100 con democrático.
Que haya cierto efecto del tipo de docente (antecedente) sobre el rendimiento de los alumnos (consecuente) no es equivalente a que el tipo de docente sea "la causa" del rendimiento alto o bajo. Así, vemos que hay algunos alumnos con docente autoritario que obtuvieron altos rendimientos (aunque solo sea el 6% del total) y otros que aun con docente democrático, tuvieron rendimiento bajo. Solo podemos afirmar que si el docente es democrático, hay más posibilidades de que los alumnos tengan mejor rendimiento, o que el tipo de docente es uno de los muchos factores que pueden incidir en el rendimiento de los alumnos.
Al comparar la tabla 6 con la 5 vemos que aunque ambas ofrecen frecuencias relativas, la 5 las calcula respecto del total general (los 400 casos) mientras que en la 6 los totales son las frecuencias marginales de cada fila. Son dos resultados muy diferentes, en efecto, la frecuencia de la celda 1,1 en la tabla 5 es 0,01 e indica que el 1% **del total de alumnos** tuvo docente autoritario y tuvo rendimiento alto, La misma celda en la tabla 6 tiene frecuencia de 0,06 y se lee: **de los alumnos con docente autoritario**, el 6% tuvo rendimiento alto. Veamos que la primera de las frecuencias (relativa al total) contiene información simultanea sobre las dos variables, mientras que la segunda (relativa a los totales de fila), fija una categoría para una de las variables: al hablar de los alumnos con docente autoritario, estamos restringiendo el conjunto completo, ya no es un juicio sobre los 400 alumnos del total, sino solo sobre los 90 que cumplen con ese requisito (haber tenido docente autoritario). La frecuencia está condicionada por ese requisito. Esto equivale a decir que las frecuencias relativas cambian cuando se establece una condición como la mencionada. En general, la proporción de alumnos con alto rendimiento es del 66%, pero si agregamos el dato que indica que estamos tratando solo con los que tuvieron docente autoritario, entonces la proporción decae al 6%. El tipo de docente es una condición que imponemos cuando calculamos estas frecuencias relativas a los totales de fila. Volveremos sobre este tema más adelante.

El modo en que se calcularon las frecuencias relativas (o los porcentajes) en la tabla 6 fue tomando como total al número de alumnos con cada tipo de docente. No es ésta la única opción posible, ¿por qué no lo hicimos dividiendo por los totales de cada nivel de rendimiento? Es decir ¿qué hizo que eligiéramos en esta tabla las filas

y no las columnas como totales para el cálculo de los porcentajes? En los casos en que tratemos con relaciones asimétricas, como lo es el del ejemplo, siempre elegiremos como denominador a los totales de la variable antecedente, porque queremos ver qué diferencia hay entre los grupos que definen sus categorías. En nuestro caso, el interés se centra en saber si el cambio de docente autoritario a democrático implica diferencia en el rendimiento de los alumnos. No es importante si la variable antecedente se ubica en las filas o en las columnas, son sus totales los que usaremos para el cálculo de los porcentajes.

Con el paso de las frecuencias simples a las relativas, hemos avanzado en la detección de la relación entre las dos variables, pero aun no podemos cuantificar su intensidad. Para ello existe una gran cantidad de coeficientes que se usan para reconocer si se trata de relaciones fuertes, débiles o simplemente inexistentes. Estos coeficientes varían según el nivel de medición de las variables, según el número de categorías, la simetría de la relación y, en especial, en el aspecto que analizan de la relación y el modo en que se interpretan. En este capítulo solo nos detendremos en uno de ellos que es de utilidad para tratar relaciones entre variables que tienen dos categorías cada una, es decir entre dos variables dicotómicas. En los próximos capítulos trataremos con otros coeficientes.
El coeficiente que usaremos para evaluar la intensidad de una relación entre dos variables dicotómicas se denomina *Q de Kendall* y en su cálculo tiene en cuenta el modo en que las frecuencias se distribuyen entre las cuatro celdas de la tabla.
Para ejemplificar el uso de este coeficiente, transformaremos la relación del caso anterior, dejando de lado a los alumnos de rendimiento medio, con lo que la tabla queda:

Tabla 7: Alumnos primarios por rendimiento según tipo de docente

Tipo de docente	Rendimiento		
	Alto	Bajo	Total
Democrático	260	10	270
Autoritario	5	50	55
Total	265	60	325

La concentración de la mayoría de los casos en las celdas de una de las diagonales de la tabla es una señal de la asociación existente entre las dos variables. El coeficiente Q se calcula operando con esas frecuencias del siguiente modo:

$$Q = \frac{260 * 50 - 5 * 10}{260 * 50 + 5 * 10} = \frac{12950}{13050} = 0{,}992$$

En el numerador, hemos multiplicado las frecuencias de una de las diagonales (260*50) y le hemos restado el producto de las frecuencias de la otra diagonal (5*10). En el denominador, hemos sumado los mismos dos productos. De manera simbólica, si representamos a las frecuencias de las celdas como A, B, C y D, tenemos

A	B
C	D

$$Q = \frac{A * D - C * B}{A * D + C * B}$$

El cálculo de este coeficiente da un número que puede ser positivo o negativo pero que siempre se encuentra entre –1 y 1.

$$-1 \leq Q \leq 1$$

En este coeficiente, el signo no tiene interpretación, se consideran iguales, por ejemplo, los valores 0,80 y -0,80. Esto se debe a que el signo que resulta depende del orden (arbitrariamente elegido) con que se hayan dispuesto las filas y las columnas. Si en la tabla 7 cambiamos las filas:

Tabla 7 bis: Alumnos primarios por rendimiento según tipo de docente

Tipo de docente	Rendimiento		
	Alto	Bajo	Total
Autoritario	5	50	55
Democrático	260	10	270
Total	265	60	325

La tabla es la misma, pero el coeficiente cambia de signo:

$$Q = \frac{5 * 10 - 260 * 50}{5 * 10 + 260 * 50} = \frac{-12950}{13050} = -0{,}992$$

Este resultado se interpreta igual que si fuera positivo.

Cuanto más próximo a uno (1) o a menos uno (-1) sea el coeficiente, tanto más intensa es la relación entre las dos variables. Los valores del coeficiente cercanos a cero indican una relación entre las variables que es débil o inexistente.

Por lo tanto, el valor obtenido en el ejemplo anterior señala una relación intensa entre las dos variables, pudiendo llevarnos a afirmar que el tipo de docente tiene, según estos datos, mucho que ver con los resultados que los alumnos obtienen. Dicho de otra manera, entre los muchos factores que pueden incidir en el rendimiento de los alumnos, el tipo de docente juega un papel importante.
En el caso extremo que el coeficiente sea igual a uno (o a menos uno) diremos que la relación es perfecta. Se trata de un caso ideal, no factible de ser observado en la realidad, pero que sirve para establecer el valor límite del coeficiente. Un ejemplo en que esto sucedería es si las frecuencias de la tabla anterior fueran como las siguientes:

Tabla 8: Alumnos primarios por rendimiento según tipo de docente

Tipo de docente	Rendimiento		
	Alto	Bajo	Total
Democrático	270	0	270
Autoritario	0	55	55
Total	270	55	325

Aquí resulta que todos los alumnos que tuvieron docentes autoritarios tienen bajo rendimiento y todos los que tuvieron docentes democráticos, alto rendimiento. En esta tabla

$$Q = \frac{270 * 55 - 0 * 0}{270 * 55 + 0 * 0} = \frac{14850}{14850} = 1$$

El valor 1 se interpreta entonces indicando que la relación entre el tipo de docente y el rendimiento de los alumnos es perfecta.

El caso contrario es aquél en el que no haya relación alguna entre las variables, allí es cuando el coeficiente alcanza (en valor absoluto) su mínimo valor posible, cero. Otra vez se trata de un caso ideal, porque muy improbablemente se encontrarán en la realidad observaciones que lleven a un coeficiente que sea exactamente cero.
Modifiquemos nuevamente las frecuencias de nuestra tabla para ejemplificar esa situación ficticia:

Tabla 9: Alumnos primarios por rendimiento según tipo de docente

Tipo de docente	Rendimiento		
	Alto	Bajo	Total
Democrático	54	216	270
Autoritario	11	44	55
Total	65	260	325

En este caso los alumnos están distribuidos en las celdas del mismo modo si se trata de docentes autoritarios o democráticos, lo cual puede verse mejor si se calculan las frecuencias relativas por filas, como hicimos antes:

Tabla 10: Alumnos primarios por rendimiento según tipo de docente

	rendimiento		
tipo de docente	alto	bajo	total
democrático	0,20	0,80	1,00
autoritario	0,20	0,80	1,00
total	0,20	0,80	1,00

Aquí encontramos que un 20% de los alumnos tiene rendimiento alto y un 80% bajo, ya sea que hayan tenido docentes autoritarios o democráticos.
El cálculo del coeficiente da ahora:

$$Q = \frac{54 * 44 - 11 * 216}{54 * 44 + 11 * 216} = \frac{0}{4752} = 0$$

Este valor indica que no hay relación entre las variables, es decir que, según estos datos, el tipo de docente no hace diferencia alguna en el rendimiento de los alumnos.

El **coeficiente Q de Kendall** mide la intensidad de la relación entre dos variables dicotómicas comparando la concentración de frecuencias en las diagonales. Alcanza su valor máximo cuando todos los casos se ubican sobre una diagonal y la relación es perfecta. Alcanza su mínimo valor cuando las frecuencias están distribuidas de manera proporcional entre las celdas y las variables son independientes.

Una limitación importante de este coeficiente aparece cuando la distribución de las frecuencias es “rinconal”. Esto quiere decir que

una de las frecuencias es cero, como sucedería si, en la tabla 8, no se hubiesen encontrado docentes democráticos con alumnos de bajo rendimiento:

Tabla 11: Ejemplo de distribución rinconal, alumnos primarios por rendimiento según tipo de docente

Tipo de docente	Rendimiento		
	Alto	Bajo	Total
Democrático	54	0	54
Autoritario	11	44	55
Total	65	44	109

En este ejemplo —y siempre que una celda tenga frecuencia cero—, el coeficiente Q dará valor 1 (ó -1) y esto no debe interpretarse como una asociación perfecta.

Terminaremos esta introducción a la relación entre variables con una referencia al problema de la causalidad. El hecho de haber encontrado que, en una relación asimétrica, existe una asociación intensa entre las variables, no nos lleva inmediatamente a suponer que la antecedente sea causa del consecuente. En toda explicación de un fenómeno, en especial de los fenómenos sociales, la causalidad es múltiple, es casi siempre imposible atribuir una causa única a la explicación de un hecho. Desde el sentido común es frecuente enunciar que "todo tiene una causa", pero en el dominio de la investigación en ciencias sociales los hechos que nos interesa analizar tienen múltiples causas, las cuales compiten entre sí en la explicación. Por lo tanto, descubriendo relaciones entre variables podemos aportar a la inclusión o exclusión de variables como factores explicativos de un fenómeno dado, pero no a "determinar su causa". Podremos afirmar qué factores hacen más probable la aparición de un fenómeno dado, bajo qué condiciones su ocurrencia es más frecuente o inclusive indicar cuáles son las variables más importantes para que el fenómeno suceda, pero muy difícilmente lleguemos a afirmaciones del tipo *X es la causa de Y*.
Pensemos por ejemplo en fenómenos psicosociales complejos, como la delincuencia juvenil. El tipo de hogar del que las personas provienen puede tener efecto, la relación con los padres, el abandono temprano de la escuela, la estructura familiar actual, y pueden seguir enumerándose factores que contribuirían a explicar que algunas personas desarrollen conductas delictivas y otras no. Pero no será

posible alcanzar una explicación completa del fenómeno, en una expresión ingenua como *la causa de la delincuencia es...*

El concepto de independencia estadística

Formulemos ahora el problema de manera inversa, interrogándonos por las condiciones en que puede decirse que dos variables son independientes. Intuitivamente la independencia entre dos eventos puede hacerse equivalente al hecho que la ocurrencia de una de ellas no tiene efecto en la de la otra. Así, las oportunidades que un evento ocurra serán iguales tanto si el otro evento sucedió como si no lo hizo. Cuando decimos que X no tiene efectos sobre Y, indicamos que Y sucede tanto si X está presente como si no lo está. La independencia de dos variables es equivalente a que no haya asociación entre ellas.
Repitamos el cruce de las variables tipo de docente y rendimiento, ahora con frecuencias diferentes.

Tabla 12: Alumnos primarios por rendimiento según tipo de docente, frecuencias absolutas

tipo de docente	rendimiento		
	alto	bajo	total
democrático	170	10	180
autoritario	30	90	120
total	200	100	300

Tabla 13: Alumnos primarios por rendimiento según tipo de docente, frecuencias relativas por filas

tipo de docente	rendimiento		
	alto	bajo	total
democrático	0,94	0,06	1,00
autoritario	0,25	0,75	1,00
total	0,67	0,33	1,00

En la última fila, las frecuencias marginales indican que en la muestra observada hubo un 67% de alumnos de rendimiento alto y 33% de rendimiento bajo, sin tener en cuenta el tipo de docente. Este dato nada dice sobre la relación entre las variables sino que proviene del modo en que resultó la composición de la muestra.
Si el tipo de docente no tuviera efecto en el rendimiento, esperaríamos que haya igual proporción de alumnos con rendimiento alto y bajo entre docentes de diferente tipo. Si del total de alumnos, el 67% tiene

rendimiento alto, los docentes democráticos deberían tener un 67% de alumnos con rendimiento alto y también debería ser así para los docentes autoritarios. De modo que, de los 180 alumnos que tuvieron docente democrático, 120 (que constituyen aproximadamente el 67% de 180) deberían haber tenido rendimiento alto. Análogamente, el 67% de 120, (aproximadamente 80 alumnos) son los que debería haber con alto rendimiento y docente autoritario. Las frecuencias de las celdas de los alumnos con bajo rendimiento se obtienen usando ahora el porcentaje del 33% sobre los mismos totales (180 y 120). Puede entonces construirse una nueva tabla con las frecuencias que se esperaría encontrar si las dos variables fueran independientes, es decir si el tipo de docente no tuviera efecto alguno en el rendimiento de los alumnos.

Tabla 14: Frecuencias esperadas bajo la hipótesis de independencia correspondiente a la tabla 12

	Alto	Bajo	Total
Democrático	120	60	180
Autoritario	80	40	120
Total	200	100	300

Observemos algunos detalles de esta tabla. En primer lugar, las frecuencias marginales no han cambiado, los totales son los mismos y solo se trata de un reordenamiento de las frecuencias conjuntas bajo la hipótesis de independencia de las dos variables[32].
Tratemos ahora de formalizar las operaciones que condujeron a esta segunda tabla. Los valores 67% y 33% provienen de las proporciones de casos en cada una de las categorías de la variable "rendimiento de los alumnos", y se calcularon como $\frac{200}{300}$ y $\frac{100}{300}$ respectivamente. Luego esas proporciones se multiplicaron por los totales de casos de cada categoría de la variable "tipo de docente". De esa operación obtuvimos 120 como 180*0,66, que daría lo mismo como $180 * \frac{200}{300}$.
El valor 80 proviene de 120*0,66 o bien de $120 * \frac{200}{300}$.

[32] Como consecuencia de ello, de las cuatro celdas solo es necesario calcular una frecuencia, ya que las demás pueden obtenerse restando de los totales de filas y de columnas. Una vez que sabemos que la frecuencia de la celda 1,1 es 120, podemos obtener 60 como lo que resta para llegar a 180 (de la primera fila), 80 como la diferencia con 200 (de la primera columna) y 40 como lo que le falta a 60 para llegar a 100 (segunda columna) o lo que le falta a 80 para alcanzar 120 (segunda fila).

60 es $180 * \frac{100}{300}$.
Finalmente, 40 resulta de hacer $120 * \frac{100}{300}$.

De manera general entonces, obtenemos cada una de las frecuencias de la segunda tabla multiplicando la frecuencia marginal de su fila por la de su columna y dividiendo por el total general. En símbolos:

$$f_{ij} = \frac{f_i * f_j}{n}$$

Si las dos variables fueran independientes (con más precisión se dice estadísticamente independientes), las frecuencias conjuntas serían como las que calculamos con este procedimiento. ¿Y qué sería en ese caso de las frecuencias relativas? Dividiendo todo por el total obtenemos:

Tabla 15: Frecuencias esperadas bajo la hipótesis de independencia correspondiente a la tabla 12

	Alto	Bajo	Total
Democrático	0,40	0,20	0,60
Autoritario	0,27	0,13	0,40
Total	0,67	0,33	1,00

Puede llegarse directamente a las frecuencias relativas, porque la frecuencia absoluta de cada celda es:

$$f_{ij} = \frac{f_i * f_j}{n}$$

Y relativa de esa celda:

$$f'_{ij} = \frac{f_{ij}}{n}$$

Si reemplazamos, nos queda:

$$f'_{ij} = \frac{f_{ij}}{n} = \frac{\left(\frac{f_i * f_j}{n}\right)}{n} = \frac{f_i}{n} * \frac{f_j}{n} = f'_i * f'_j$$

Más brevemente:

$$f'_{ij} = f'_i * f'_j$$

Es decir que, si las variables fueran independientes, cada frecuencia relativa será producto de las correspondientes frecuencias relativas

marginales. Ahora podemos dar una definición de independencia estadística.

Dos variables son **estadísticamente independientes** si la frecuencia relativa de cada celda es igual al producto de las frecuencias relativas marginales de la fila y la columna a las que la celda pertenece.

En efecto, cada frecuencia conjunta de la tabla 14 es el producto de las marginales correspondientes. Verifique que es así.

En este capítulo solo hemos tratado con variables nominales, y en un caso también ordinales, como en el Esquema 6, pero nada hemos dicho aun de las variables intervalares y proporcionales. En el capítulo 2 vimos que una tabla de distribución de frecuencias no puede listar todas las categorías de una variable de estos niveles, sino que deben construirse intervalos de valores. Eso mismo puede hacerse para construir una tabla bivariada, como las que vimos en este capítulo para variables intervalares y proporcionales. De ese modo obtendríamos una tabla como:

Esquema 7: Disposición de las variables para analizar la relación entre los años de escolarización y el ingreso mensual individual

		Ingresos mensuales					Total
		<1000	1000 a 2000	2000 a 3000	3000 a 4000	> 4000	
Años de escolarización	< 5						
	5-10						
	10-15						
	> 15						
Total							

Pero en el próximo capítulo veremos que para variables de estos niveles de medición se cuenta con procedimientos más simples y más eficaces que permiten analizar con más detalle sus relaciones.

Actividad práctica de repaso 4

1. En una investigación realizada por la Cátedra de Psicología Evolutiva de la Adolescencia y dirigida por la Lic. Cardozo (2009), se observó que los adolescentes que poseen mayor ansiedad social presentan mayores conductas de retraimiento y menores conductas de liderazgo.

a. La relación entre ansiedad social y conductas de retraimiento es:
Simétrica Asimétrica Directa Inversa

b. La relación entre ansiedad social y conductas de liderazgo es:
Simétrica Asimétrica Directa Inversa

2. En una investigación realizada desde la Secretaría de Bienestar Estudiantil de la Universidad Nacional de Córdoba se indagó si existía asociación entre la situación laboral de los estudiantes y su permanencia o abandono de la carrera. Para ello se trabajó con una muestra de 250 estudiantes y se observó que, de 60 que trabajan 20 abandonaron sus estudios y que, de los que no trabajan 80 abandonaron.
a. Indique cuáles son las variables cuya relación se analiza
b. ¿Cuáles son las categorías de cada una de las variables?
c. ¿Se podría plantear como una relación simétrica o asimétrica?, en el segundo caso, indique qué variable podría ser antecedente y cuál consecuente.
d. Elabore una tabla de contingencia donde se vislumbre la relación entre las variables presentadas (use las filas y las columnas que sean necesarias).
e. Determine la dimensión de la tabla resultante.
f. Complete, en esa tabla, las casillas con las frecuencias marginales y conjuntas.
g. Evalúe la intensidad de la asociación

3. En un estudio dirigido a comparar el rendimiento de alumnos según su lugar de procedencia, se obtuvo la siguiente tabla:

	Condición al cabo del cursado			
Lugar de procedencia	Promocionados	Regulares	Libres	Total
Esta ciudad	150	200	50	400
Otra ciudad (de Argentina)	100	120	30	250
Otro país	30	40	5	75
Total	280	360	85	725

a. Indicar las categorías de las variables de filas y de columnas
b. ¿Cuál es la dimensión de la tabla?
c. Calcule la frecuencia relativa de la celda 2,3
d. ¿Qué proporción de alumnos de esta ciudad promocionó?
e. ¿y de otro país?
f. ¿Qué proporción del total de alumnos proviene de otra ciudad de Argentina?
g. ¿Qué proporción de promocionados son de esta ciudad?
h. ¿y de otro país?

Capítulo 5: Intensidad y forma de la relación entre variables

Leonardo Medrano
Eduardo Bologna

En el capítulo anterior hemos tratado la relación entre dos variables en escalas nominales, y señalamos que si se trata de variables de nivel superior es posible crear categorías y tratarlas del mismo modo. En cuanto a la medida de la intensidad de la relación, nos hemos limitado al caso de dos variables dicotómicas, es decir, con dos categorías en cada una, con lo que la tabla resultante es de dos por dos y allí es que calculamos el coeficiente Q de Kendall. Nos proponemos en este capítulo ampliar el dominio de nuestro análisis, incorporando herramientas que permitirán poner a prueba la hipotética relación entre dos variables de nivel nominal con más de dos categorías cada una y variables de nivel superior (ordinales y métricas).

Variables nominales con más de dos categorías cada una

La distancia entre frecuencias esperadas y observadas
Sobre el final del capítulo anterior presentamos el concepto de independencia estadística y vimos la manera de calcular las frecuencias de las celdas que se esperarían encontrar si las variables fueran independientes. Para hacer esto es suficiente multiplicar las frecuencias marginales correspondientes a cada celda y dividir el resultado por el total de casos. Veamos el ejemplo siguiente, que se dirige a analizar la posible relación del tipo de violencia[33] con el lugar donde sucede, a partir de una muestra de 500 casos seleccionados en tres áreas geográficas:

[33] Según la clasificación sugerida por el Informe Mundial sobre la Violencia y la Salud, Organización Panamericana de la Salud, 2003

Tabla 1: Clasificación de diferentes tipos de violencia según área donde se manifiesta

Tipo de violencia	Área			
	Ciudades grandes	Ciudades pequeñas	Áreas rurales	Total
Autoinfligida	100	35	15	150
Interpersonal	110	100	90	300
Colectiva	35	10	5	50
Total	245	145	110	500

Una primera aproximación consiste en calcular frecuencias relativas. Dado que nuestro interés está en comparar el tipo de violencia según las áreas, calcularemos los porcentajes según las columnas de la tabla 1 y resulta:

Tabla 2: Frecuencia relativas por columnas de la clasificación de diferentes tipos de violencia según área donde se manifiesta

Tipo de violencia	Área			
	Ciudades grandes	Ciudades pequeñas	Áreas rurales	Total
Autoinfligida	41%	24%	14%	30%
Interpersonal	45%	69%	82%	60%
Colectiva	14%	7%	5%	10%
Total	100%	100%	100%	100%

Si no consideramos el área, se ve que la violencia interpersonal es la más frecuente (60% del total, que se observa en la última columna), seguida de la autoinfligida con el 30%. Este patrón de distribución en las distintas formas de violencia se mantiene en las diferentes áreas, pero en más acentuado en las rurales, donde la categoría modal (que sigue siendo interpersonal) alcanza el 82% del total del área. Por el contrario, la violencia autoinfligida, que es el 30% del total, sube al 41% en grandes ciudades y solo representa el 14% de las formas de violencia que se observan en áreas rurales. Así, parecería que hay diferencia en la distribución de los tipos de violencia según las áreas que están considerándose.

Codificando como 1, 2 y 3 las áreas y del mismo modo el tipo de violencia, el programa InfoStat® muestra las frecuencias absolutas y las relativas por columnas así:

Tablas de contingencia

Frecuencias absolutas
En columnas: área

tipo violencia	1	2	3	Total
1	100	35	15	150
2	110	100	90	300
3	35	10	5	50
Total	245	145	110	500

Frecuencias relativas por columnas
En columnas: área

tipo violencia	1	2	3	Total
1	0,41	0,24	0,14	0,30
2	0,45	0,69	0,82	0,60
3	0,14	0,07	0,05	0,10
Total	1,00	1,00	1,00	1,00

Buscaremos ahora de cuantificar la intensidad de esa relación, para lo que nos preguntaremos cuáles serían las frecuencias de las celdas si el tipo de violencia fuera independiente del área donde sucede, es decir, si se observara la misma proporción de los distintos tipos de violencia en todas las áreas. Usemos el concepto de independencia estadística para calcular las frecuencias esperadas correspondientes a la tabla 1

Tabla 3: Frecuencias esperadas bajo la hipótesis de independencia entre el tipo de violencia y el área donde se observa

Tipo de violencia	Área			
	Ciudades grandes	Ciudades pequeñas	Áreas rurales	Total
Autoinfligida	74	43	33	150
Interpersonal	147	87	66	300
Colectiva	24	15	11	50
Total	245	145	110	500

Estas frecuencias están calculadas como indicamos en el capítulo 5, haciendo

$$f_{ij} = \frac{f_i * f_j}{n}$$

Por ejemplo, la frecuencia de la celda 1,1 resultó de $f_{11} = \frac{150 * 245}{500} = 73,5$, que redondeamos a 74.

Solicitada a InfoStat®, esta salida tiene la forma:

Tablas de contingencia[34]

Frecuencias absolutas
En columnas: área
tipo

violencia	1	2	3	Total
1	100	35	15	150
2	110	100	90	300
3	35	10	5	50
Total	245	145	110	500

Frecuencias esperadas
En columnas: área
tipo

violencia	1	2	3	Total
1	73,5	43,5	33,0	150,0
2	147,0	87,0	66,0	300,0
3	24,5	14,5	11,0	50,0
Total	245,0	145,0	110,0	500,0

Estas últimas son las frecuencias que esperaríamos encontrar si no hubiera relación entre las variables, si fueran independientes. A ellas debemos compararlas con las que realmente hemos encontrado, las que se denominan frecuencias observadas.

Si halláramos que nuestras frecuencias observadas son muy similares a las esperadas bajo la hipótesis de independencia, diríamos que las variables "están cerca" de ser independientes, o lo que es equivalente, que había escasa relación entre ellas. Por el contrario, si las frecuencias observadas son muy diferentes de las esperadas, creeríamos que las variables "están lejos" de ser independientes, es decir, que habría alguna relación entre ellas. Para decidir, debemos comparar la tabla 1 con la 3. Para simplificar quitamos los nombres de las categorías y las frecuencias marginales, así, las que debemos comparar son las siguientes frecuencias:

[34] En la salida no están redondeados los decimales de las frecuencias esperadas

Tabla 4: Frecuencias observadas y esperadas para la tabla 1

Frecuencias

Observadas		
100	35	15
110	100	90
35	10	5

Esperadas		
74	43	33
147	87	66
24	15	11

Una opción para medir la distancia entre los dos conjuntos de frecuencias es la de restar las correspondientes de cada celda; pero si hacemos eso nos encontraremos con un problema parecido al que tuvimos cuando intentamos observar la dispersión restando los valores de la media: la suma nos da cero. Por única vez realizaremos esta operación de manera manual:

$$(100-74)+(35-43)+(15-33)+(110-147)+(100-87)$$
$$+(90-66)+(35-24)+(10-15)+(5-11)=0$$

Obtenemos este resultado porque las frecuencias marginales son fijas y lo que una celda tiene de más, lo tiene otra de menos. Siempre sucederá así y por esa razón, no podemos saber si las observadas están cerca o lejos de las esperadas con el procedimiento directo de restarlas. Por el contrario, para medir la distancia entre los dos conjuntos de frecuencias (observadas y esperadas) se usa la siguiente expresión:

$$\sum_{\substack{i=1\\j=1}}^{\substack{i=f\\j=c}} \frac{(f_{ij}^{o}-f_{ij}^{e})^2}{f_{ij}^{e}}$$

La expresión nos dice que deben restarse cada una de las frecuencias esperadas de cada observada correspondiente, elevar esa diferencia al cuadrado[35] y dividir el resultado por cada una de las frecuencias esperadas. Los subíndices mantienen la notación del capítulo anterior: i es el índice de filas, que va desde la primera (i=1) hasta la última (f es el número total de filas); j es el índice de las columnas, que también empieza en 1 (j=1) y termina en c, que es el número total de columnas[36]. Vamos a aplicarla una vez, solo para ver su funcionamiento, luego la pediremos al programa:

[35] El mismo recurso que se usó cuando se definió la varianza y no era posible usar la suma de los desvíos porque daba cero. Nuevamente aquí, usamos el exponente 2 para volver positivos a los números negativos.

[36] Recordemos que, de manera general la dimensión de la tabla es *f X c*, filas por columnas.

$$\frac{(100-74)^2}{74}+\frac{(35-43)^2}{43}+\frac{(15-33)^2}{33}+\frac{(110-147)^2}{147}+\frac{(100-87)^2}{87}$$
$$+\frac{(90-66)^2}{66}+\frac{(35-24)^2}{24}+\frac{(10-15)^2}{15}+\frac{(5-11)^2}{11}=50{,}40$$

El número que resulta de esta operación se llama puntaje chi cuadrado (o también ji cuadrado), se indica con el símbolo χ^2 y es una medida de la distancia a la que se encuentran las frecuencias observadas de las que se esperaría encontrar si las variables fueran independientes.

Cuando se pide a InfoStat®, se obtiene:

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	**50,19**	4	<0,0001
Chi Cuadrado MV-G2	52,25	4	<0,0001
Coef.Conting.Cramer	0,18		
Coef.Conting.Pearson	0,30		

De esta salida por ahora solo nos interesa lo que hemos destacado (Chi Cuadrado Pearson); más adelante usaremos otros de los resultados de la salida. El valor no es idéntico al que hayamos manualmente (50,19 contra 50,40), eso se debe a que habíamos redondeado las frecuencias esperadas y el programa trabaja con los decimales, pero la diferencia no es importante.
Este número, el puntaje χ^2 no puede ser negativo, ya que proviene de la suma de números elevados al cuadrado. Solo puede ser cero si todos los términos de la suma son cero, es decir, si cada frecuencia observada es exactamente igual a la esperada correspondiente. En ese caso no habría duda en decir que las variables son independientes, cumplirían exactamente con la definición de independencia estadística. El puntaje χ^2 indica si las frecuencias observadas están cerca o lejos de las esperadas, pero ¿qué tan grande debe ser para que consideremos lejanas a las frecuencias?
Dos problemas que tiene este puntaje son: que puede ser indefinidamente grande y que su valor depende del número de casos que se evalúan y de la dimensión de la tabla. Así, por ejemplo, si multiplicamos por 10 todas las frecuencias de la tabla 3, obtenemos:

Tabla 5: Ejemplo de expansión artificial del total de casos de la tabla 1

Tipo de violencia	Área			
	Ciudades grandes	Ciudades pequeñas	Áreas rurales	Total
Autoinfligida	1000	350	150	1500
Interpersonal	1100	1000	900	3000
Colectiva	350	100	50	500
Total	2450	1450	1100	5000

Aunque los valores absolutos son diez veces más grandes, no hubo cambios en las frecuencias relativas, por ejemplo, en la celda 1,1: $\frac{1000}{2450} = 0{,}41\,(41\%)$ lo mismo que había dado esa celda en la tabla 2. Cualquiera sea la intensidad de la relación entre estas dos variables, ésta no ha cambiado porque hayamos multiplicado todo por 10, sin embargo, si calculamos el puntaje χ^2 en la tabla 5 obtenemos 501,9, es decir un número 10 veces más grande. Entonces el puntaje χ^2 puede cambiar muy ampliamente sin que cambien las frecuencias relativas; esto nos dice que ese puntaje no mide la intensidad de la relación. Para hacerlo debemos eliminar el efecto de la cantidad de casos y también de la dimensión de la tabla. Calcularemos dos coeficientes que nos permitan evaluar el grado o intensidad de la relación y que tengan un límite superior de modo que podamos juzgarlos como elevados o bajos. El primero de ellos es el **coeficiente de contingencia**, C de Pearson, se calcula del siguiente modo a partir del puntaje χ^2 obtenido antes:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Este coeficiente no puede ser menor que cero (0) y solo toma ese valor si las variables son independientes (es decir cuando $\chi^2 = 0$). Tampoco puede ser mayor que uno (1), pero su valor máximo depende de la dimensión de la tabla.
Solo en el caso particular en que la tabla sea cuadrada (misma cantidad de filas que de columnas), el valor máximo del coeficiente es:

$$C_{max} = \sqrt{\frac{f-1}{f}} \qquad \text{o bien} \qquad C_{max} = \sqrt{\frac{c-1}{c}}$$

Ya que nos referimos a tablas cuadradas, en las que $f = c$.

De este modo alcanzamos un coeficiente que nos indica el grado de la asociación entre dos variables que mejora lo que habíamos

logrado con el Q de Kendall, ya que este es apto para tablas de cualquier dimensión, no solo para las de 2 X 2. Reemplacemos los valores para el ejemplo anterior:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{50{,}19}{50{,}19 + 500}} = 0{,}30$$

Para decidir si este resultado es alto o bajo, es decir, si la relación es fuerte o débil, calculemos el máximo que podría haber alcanzado para una tabla de 3X3:

$$C_{max} = \sqrt{\frac{f-1}{f}} = \sqrt{\frac{3-1}{3}} = \sqrt{\frac{2}{3}} = 0{,}82$$

Entonces el valor que hemos encontrado es moderado, y nos indica que la relación entre el tipo de violencia y el tamaño de las ciudades no es intensa.

El segundo (y último) coeficiente que calcularemos para variables nominales está también basado en el puntaje χ^2 y tiene un valor máximo de 1 (uno). Se llama **coeficiente V de Cramer** y se calcula así:

$$V = \sqrt{\frac{\chi^2}{n * \min(f-1; c-1)}}$$

La expresión *min(f-1, c-1)*, a pesar de su misterioso aspecto, tiene un significado sencillo: se refiere al mínimo entre el número de filas menos uno y el número de columnas menos uno. Para obtenerlo se resta 1 al número de filas, luego se resta 1 al número de columnas y se elije el menor de los dos. Si la tabla es de 3X2 se hace 3 - 1=2 y 2 - 1=1, entre 2 y 1 el mínimo es 1 y ése es el número que ubicamos en el denominador, multiplicando a n. En nuestro ejemplo, las filas y las columnas son tres, por lo que hacemos el mínimo entre 3-1 y 3-1, el resultado es 2, y obtenemos para el coeficiente de Cramer:

$$V = \sqrt{\frac{\chi^2}{n * \min(f-1, c-1)}} = \sqrt{\frac{50{,}19}{500 * \min(3-1, 3-1)}} = \sqrt{\frac{50{,}19}{500 * 2}} = 0{,}45$$

Como el valor máximo que puede alcanzar este coeficiente es 1 (uno), diremos que se trata de una relación moderada entre las dos variables.

Variables de nivel ordinal

Si nuestro problema es el de describir la relación entre variables cuyas categorías están ordenadas, es decir variables ordinales, los coeficientes anteriores son válidos: pero como sucedió con las medidas descriptivas, el mayor nivel de medición permite calcular coeficientes más elaborados y que, por esa razón, nos informen más acerca de las variables que se analizan. Un punto a tener en cuenta es que cuando se trata con dos variables, la del menor nivel de medición es la que manda. Así, para relacionar una ordinal y una nominal, debe usarse un coeficiente V o C, como si fueran las dos nominales.

Si las dos variables son ordinales o si una es ordinal y la otra intervalar o proporcional, es posible calcular un coeficiente que tiene en cuenta los "rangos" es decir la posición de cada categoría respecto de las demás, su carácter de primera, segunda, etc., en suma, el orden. Sea el problema de indagar por la relación que podría haber entre el resultado que obtienen los alumnos al rendir un examen de ingreso a una carrera y el nivel de educación de sus madres, consideraremos como variables *el nivel máximo de educación de la madre* de cada uno y *el orden de mérito alcanzado en el ingreso* a esa carrera; ambas variables son ordinales. Tratemos solo la situación de pocos casos por ahora. Si para el orden de mérito codificamos como 1, 2, 3… el primer lugar en el ingreso, el segundo, etc. y para la educación de la madre usamos 1 = universitario completo, 2 = universitario incompleto, etc., entonces el fragmento de la matriz de datos para 8 observaciones tendría una forma como esta:

Sujeto	Orden de mérito en el ingreso	Educación de la madre
1	1	2
2	2	2
3	3	3
4	4	2
5	5	3
6	6	4
7	7	5
8	8	3

A la que hemos ordenado según los valores de la primera variable (orden de mérito). A estos datos no conviene presentarlos en una tabla de doble entrada, porque cada orden de mérito corresponde a un único individuo, por lo que obtendríamos una tabla tan poco resumida como la siguiente:

Tabla 6: Distribución conjunta de las frecuencias del orden de mérito en el ingreso a una carrera universitaria y el nivel de educación de la madre.

		Educación de la madre							total
		1	2	3	4	5	6	7	
Orden de mérito en el ingreso	1	0	1	0	0	0	0	0	1
	2	0	1	0	0	0	0	0	1
	3	0	0	1	0	0	0	0	1
	4	0	1	0	0	0	0	0	1
	5	0	0	1	0	0	0	0	1
	6	0	0	0	1	0	0	0	1
	7	0	0	0	0	1	0	0	1
	8	0	0	1	0	0	0	0	1
Total		0	3	3	1	1	0	0	8

Esta tabla no es útil, ya que tiene tantas filas como la matriz de datos (porque cada orden de mérito corresponde a una sola persona), hay solo un caso en cada celda no vacía y hay muchas celdas vacías (con frecuencia cero). Por eso, cuando se trate de variables de este nivel de medición, no usaremos tablas de doble entrada para representar los datos, solo calcularemos un coeficiente que nos indique la intensidad de la relación.

Este coeficiente se llama coeficiente de **correlación por rangos, de Spearman** y para calcularlo hay que transformar los valores de las variables en rangos, de mayor a menor, de manera que al máximo valor de cada variable corresponda el 1, al siguiente el 2 y así sucesivamente. En nuestro ejemplo, el orden de mérito ya está en rangos, uno para el primero, dos para el segundo y un rango para cada persona. No es así para el nivel de educación, ya que varias personas pueden tener el mismo, a esta variable la transformaremos en rangos. El mayor nivel de educación observado es 2 (universitario incompleto) a ese valor le correspondería el rango 1 (uno), pero hay tres madres con ese nivel de educación, ellas deberían llevar los rangos 1, 2 y 3, como están empatadas, les asignamos a todas el promedio de los tres rangos: 2. Luego sigue el nivel de educación 3 (secundario completo) a quien deberíamos asignar el rango 4, pero acá también hay empate entre tres casos, corresponderían los rangos 4, 5 y 6, nuevamente usamos el promedio de los tres rangos para asignar a los tres el mismo: 5. Sigue el nivel 4 (secundario incompleto), al que asignamos el rango siguiente: 7, ya que hay solo un caso aquí; y lo mismo pasa con el nivel 5 (primario completo) al que le toca rango 8.

Resumiendo entonces, la transformación de los valores de las variables en rangos resulta así:

Tabla 7: Transformación de las categorías de la variable a rangos

Orden de mérito en el ingreso	Rango del orden de mérito
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8

Educación de la madre	Rango de la educación de la madre
2	2
2	2
3	5
2	2
3	5
4	7
5	8
3	5

Vemos que no ha sido necesario transformar los valores del orden de mérito, porque ya correspondían uno a cada sujeto.
Una vez que disponemos de los rangos, vamos a observar, para cada caso la diferencia entre el rango de una variable y de la otra, esas diferencias se llamarán *d*.

Tabla 8: Cálculo de las diferencias entre los rangos de dos variables ordinales

Rango del orden de mérito	Rango de la educación de la madre	*d*
1	2	-1
2	2	0
3	5	-2
4	2	2
5	5	0
6	7	-1
7	8	-1
8	5	3

Estas diferencias indican la distancia que hay entre los dos ordenamientos, si fueran ambos iguales (si el máximo de uno coincidiera con el máximo del otro y así en todas las categorías), tendríamos una asociación perfecta entre las dos variables. Por el contrario si el orden estuviese exactamente invertido (si el rango máximo de una variable coincidiera con el rango mínimo de la otra y así en las demás) la relación también sería perfecta, pero inversa.
La intensidad de la relación se mide entonces con el que hemos llamado coeficiente se Spearman, la expresión de su cálculo es la siguiente:

$$r_s = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

En la que:
d_i son las diferencias de rangos (calculadas en la última columna de la tabla de arriba) que en la fórmula van elevadas al cuadrado.
La sumatoria indica que ésta va desde la primera de las diferencias (*i=1*) hasta la última (*n*).
n es el número total de observaciones.

Este coeficiente puede ser positivo o negativo y tiene un campo de variación igual al del Q de Kendall, es decir, entre -1 y 1, es decir:

$$-1 \leq r_s \leq 1$$

Como el coeficiente de Kendall, los valores próximos a 1 ó a -1 se interpretan como propios de una asociación fuerte (intensa) y los cercanos a 0 (cero), sean positivos o negativos, corresponden a asociaciones débiles. Si un coeficiente vale 1 ó -1 diremos que la asociación es perfecta, pero eso no es algo que suceda en la realidad, del mismo modo que si el coeficiente es exactamente 0 (cero), la asociación será nula y otra vez es muy poco común que eso suceda con datos reales.
A diferencia del coeficiente Q de Kendall, ahora el signo importa: cuando es positivo da cuenta de una relación directa entre las dos variables, una relación en la que cuando una aumenta, la otra también lo hace. Si el coeficiente es negativo indica relación inversa, el crecimiento de una variable se acompaña del decrecimiento de la otra. En este nivel de medición (ordinal) podemos hacer estos juicios, podemos decir "aumenta" o "disminuye", porque las categorías están ordenadas, por esa razón podemos analizar no solo la intensidad de la relación, sino también si se trata de una relación directa o inversa.
Se trata de dos características independientes de cada relación: puede ser fuerte y directa; o fuerte e inversa; o bien débil y directa; o débil e inversa. Un error muy frecuente es creer que si el coeficiente es negativo, la relación es débil, no es así. Es débil si el coeficiente es cercano a 0 (cero), es igualmente débil si r_s=0,03 como si r_s=-0,03, el signo del coeficiente no aporta para saber si es fuerte o débil. Del mismo modo es igual de fuerte una relación en la que r_s=0,96como una en la que r_s=-0,96.
Para obtener el valor del coeficiente en nuestro ejemplo, vamos primero a calcular las $d_i{}^2$:

Rango del orden de mérito	Rango de la educación de la madre	d_i	d_i^2
1	2	-1	1
2	2	0	0
3	5	-2	4
4	2	2	4
5	5	0	0
6	7	-1	1
7	8	-1	1
8	5	3	9

Por lo que la suma de la última columna es 20. Al reemplazar los valores en la expresión de r_s, obtenemos:

$$r_s = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6 * 20}{8^3 - 8} = 1 - \frac{120}{512 - 8} = 1 - \frac{120}{504} = 0{,}76$$

Este valor de $r_s = 0{,}76$, indica una asociación intensa y positiva entre la educación de la madre y los resultados del ingreso a la universidad. Que sea positiva quiere decir que los alumnos con madres de mayor educación obtienen mejores resultados en el ingreso a esa carrera. Como ya hemos señalado, esto no quiere decir causalidad, no significa que la causa del resultado en el ingreso sea la educación de la madre. El problema de la causalidad es teórico y depende del análisis que se hace de las relaciones entre los conceptos, en este ejemplo, la educación de la madre es uno de muchos de los factores interrelacionados, que inciden sobre el resultado que obtiene el alumno. Este coeficiente (como sucede con todos los coeficientes de asociación) no revelan la causalidad sino lo frecuente que resulta que los cambios de una variable se vean acompañados de cambios en la otra.

Al solicitar este coeficiente a InfoStat®, se obtiene la siguiente salida, de la que solo nos interesa por ahora el valor del coeficiente, que está destacado:

Coeficientes de correlación

Correlacion de Spearman: coeficientes\probabilidades

	orden de mérito	educación de la madr
orden de mérito	1,00	0,04
educación de la madr	**0,76**	1,00

En esta salida la diagonal principal siempre estará constituida por unos (1s), porque es la asociación de cada variable consigo misma. El coeficiente aparece en la posición 2,1, como el que está aquí destacado. El valor que aparece opuesto a él, el 0,04 no será interpretado hasta más adelante.
Para ilustrar con otros ejemplos de relaciones entre variables ordinales, sea que si ponemos en correspondencia el ranking de temas musicales de una semana con la frecuencia con que cada tema es reproducido en la radio, esperamos hallar que la relación sea muy fuerte y directa: los temas de mayor posición en el ranking son también los que más frecuentemente se pasan en la radio. Al revés, en la relación entre el rating de los programas de televisión y el contenido cultural que ofrecen esperamos una relación también fuerte, pero ahora inversa: los de mayor rating son habitualmente lo que menos contenido cultural tienen.

Nivel intervalar o proporcional
Si tratamos con variables intervalares o proporcionales, podríamos usar los procedimientos que referimos antes para el cálculo de la intensidad de las relaciones entre variables nominales. Para ello, deberíamos construir intervalos y tratarlos como las categorías de las dos variables. De ese modo, perderíamos la información que provee una variable cuantitativa. Por ejemplo, si disponemos de un conjunto de personas adultas de las que sabemos la edad a la que cada una se casó (o unió) por primera vez y los años de escolarización, y nos interesamos por la relación entre estas dos variables, el fragmento de la matriz de datos para estas dos variables en 30 casos es:

sujeto	años escuela	edad primera unión
1	5	18
2	6	20
3	7	17
4	7	18
5	8	21
6	8	22
7	8	24
8	8	32
9	9	16
10	9	18
11	9	19
12	9	25
13	9	27
14	11	18
15	11	25

sujeto	años escuela	edad primera unión
16	11	26
17	11	27
18	11	29
19	12	25
20	12	26
21	12	27
22	12	29
23	12	30
24	12	30
25	13	29
26	15	28
27	16	30
28	16	31
29	16	33
30	17	30

Si quisiéramos construir una tabla de doble entrada para estas dos variables, el problema es aun mayor que con las ordinales, dado que tendríamos un gran número de filas y de columnas, y resultaría casi imposible de leer. Además, muchas celdas estarían vacías y habría muy pocos casos en cada una de las restantes. Los treinta datos de la matriz anterior quedarán, en una tabla de doble entrada, así:

Tabla 9: Distribución conjunta de los años de escolarización y la edad a la que se realizó la primera unión conyugal

		Edad a la primera unión																	
		16	17	18	19	20	21	22	24	25	26	27	28	29	30	31	32	33	Total
Años de escolarización	5	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
	7	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	8	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	4
	9	1	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	5
	11	0	0	1	0	0	0	0	0	1	1	1	0	1	0	0	0	0	5
	12	0	0	0	0	0	0	0	0	1	1	1	0	1	2	0	0	0	6
	13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
	15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	3
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
	total	1	1	4	1	1	1	1	1	3	2	3	1	3	4	1	1	1	30

Nuevamente, esta no es una tabla adecuada para representar nuestros datos y sería más legible con las categorías agrupadas:

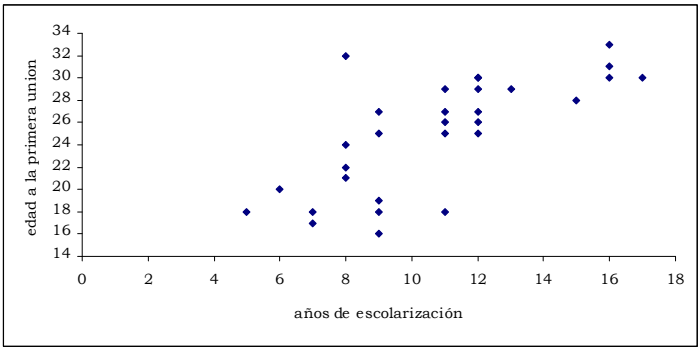
Tabla 10: Distribución conjunta de los años de escolarización y la edad a la que se realizó la primera unión conyugal (categorías agrupadas)

		Edad a la primera unión				Total
		16-20	21-25	26-30	31-35	
Años de escolarización	5-8	4	3	0	1	8
	9-12	4	3	9	0	16
	13-16	0	0	3	2	5
	17-20	0	0	1	0	1
Total		8	6	13	3	30

Ahora la tabla se lee con más facilidad y podemos tratarla como hicimos antes, como si fueran nominales y calcular un puntaje chi cuadrado y coeficientes *C de Pearson* y *V de Cramer*. Sin embargo, de ese modo vamos a perder las posibilidades de análisis que nos ofrecen las variables cuantitativas.

Para poder mantener las variables con sus verdaderos valores (sin agrupar) y tener al mismo tiempo una representación abreviada de los datos, disponemos de un recurso muy valioso: una representación gráfica de los valores que se denomina **diagrama de dispersión**.

Gráfico 1: Diagrama de dispersión de los años de escolarización y la edad a la que se realizó la primera unión conyugal

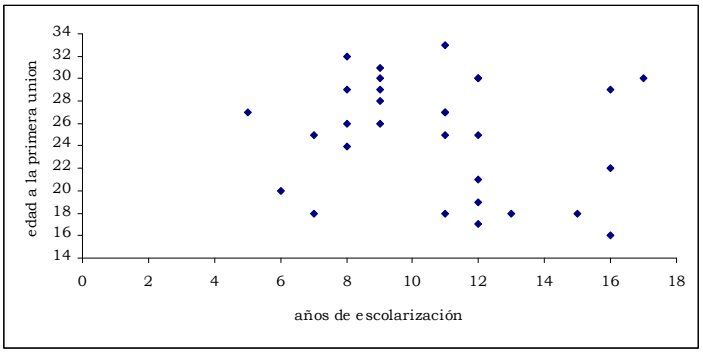


Este gráfico usa los ejes cartesianos para indicar los valores de las dos variables que estamos analizando y representa con un punto cada concordancia de dos categorías que puede corresponder a un caso o a varios. Cada punto es un par ordenado: el primer número son los años de escolarización y el segundo la edad a la que se unió por primera vez. Los ceros de la tabla 9 ya no aparecen en este diagrama. El primer punto de la izquierda del diagrama corresponde a una única persona que alcanzó 5 años de escolarización y que tuvo su primera unión a los 18 años. Observemos que los cuatro puntos alineados sobre los 8 años de escolarización, son los cuatro sujetos que aparecen como total de la fila correspondiente al 8 en la tabla 9. Lo que eran filas y columnas en todas las tablas que tratamos hasta aquí son ahora ejes coordenados, porque ya no tratamos con categorías separadas de cada variable sino con valores continuos de las variables que ahora son intervalares o proporcionales. Estos ejes se llaman **ordenadas** el vertical y **abscisas** el horizontal. En el ejemplo están representados los valores de los años de escolarización en el eje de las abscisas (primer elemento de cada par ordenado) y la edad a la primera unión en el eje de las ordenadas (segundo elemento de cada par).

La manera en que los puntos se distribuyen en el diagrama de dispersión nos da una primera aproximación a la relación entre las dos variables. Así, en el caso del ejemplo, hay una cierta tendencia creciente, en la que se vería que *globalmente*, las personas con más años de escolarización tenderían a unirse más tardíamente. Esta observación es equivalente a ver la concentración de casos en las celdas de la diagonal de una tabla bivariada, como señalamos en el capítulo 4.

Por el contrario, si nuestros datos se dispersan de este otro modo:

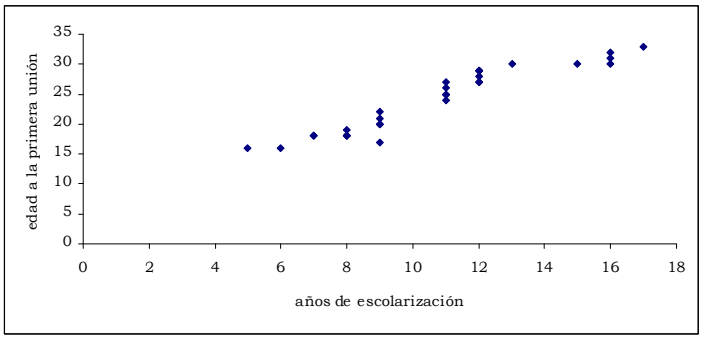
Gráfico 2: Diagrama de dispersión de los años de escolarización y la edad a la que se realizó la primera unión conyugal (relación débil)



No hay ninguna razón para creer que las variables estén relacionadas: los puntos no muestran una tendencia clara.

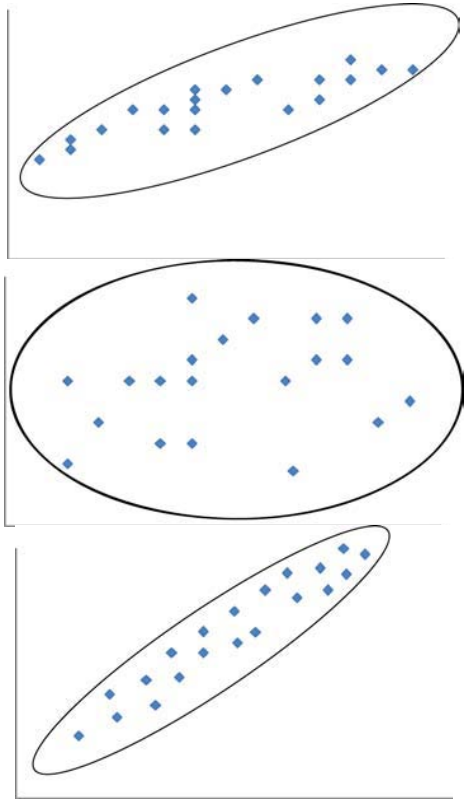
Una asociación más acentuada entre las mismas dos variables se observa en el siguiente diagrama de dispersión:

Gráfico 3: Diagrama de dispersión de los años de escolarización y la edad a la que se realizó la primera unión conyugal (relación intensa)



En el que la tendencia *lineal* es más clara, por lo que resulta más definido el efecto de la escolarización sobre la edad a la que se produce la primera unión. Decimos que la nube de puntos está más aplanada en este caso que en el anterior; en efecto, en el gráfico 2 la nube de puntos tiene forma más circular que en el 3, donde es más elíptica.

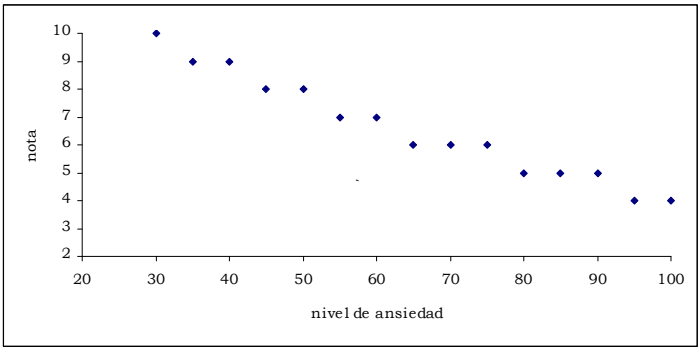
Gráfico 4: Comparación de la forma de las nubes de puntos según la intensidad de la relación



La intensidad de la relación está vinculada al achatamiento de la elipse que rodea la nube de puntos y éste al grado de alineación que los puntos tengan. Luego volveremos sobre esta idea.

El ejemplo que hemos mostrado hasta aquí corresponde a una relación directa: más años de escolarización parecen corresponder a edades más tardías para la primera unión. De manera equivalente pueden representarse relaciones inversas. Consideremos el caso de la ansiedad frente a los exámenes y la calificación que se obtiene. En un estudio realizado por el Laboratorio de Evaluación Psicológica y Educativa (LEPE) se observó que a mayor puntaje en una prueba de ansiedad ante los exámenes, menor rendimiento académico (Furlán, Cassady & Pérez, 2009). Tomando las notas obtenidas en un examen y el puntaje de la escala de ansiedad medido en una escala de 0 a 100, este resultado puede representarse gráficamente así:

Gráfico 5: Diagrama de dispersión del puntaje obtenido en una prueba de ansiedad en exámenes y las notas obtenidas

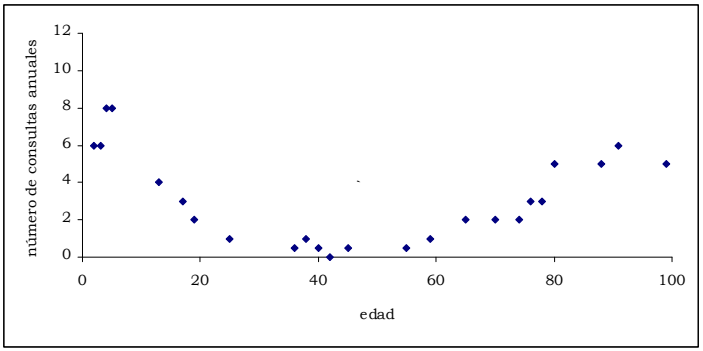


Que nos muestra que quienes llegan más ansiosos a los exámenes obtienen peores calificaciones[37]. La relación inversa queda entonces graficada como un conjunto de puntos descendentes.

Solo nos ocuparemos de relaciones como las que acabamos de ejemplificar: aquellas en las que la tendencia es creciente o decreciente, pero siempre siguiendo un camino parecido a una línea recta. Son las que llamaremos *relaciones lineales*. No son la únicas que existen; solo a modo de ilustración, veamos cómo se representa la relación entre la edad de las personas y la frecuencia con que consultan al médico. Estas dos variables son tales que, en términos muy generales y sin considerar situaciones específicas, para valores pequeños de la primera (en la infancia) las consultas son frecuentes, luego se reducen durante la adultez para volver a incrementarse en la vejez. Por eso el gráfico que las representa tiene la siguiente forma:

[37] Nuevamente aquí recordemos que esto no quiere decir causalidad, de ningún modo podemos afirmar que la ansiedad "causa" bajos resultados. Bien podría ser que los alumnos lleguen más ansiosos cuanto menos han estudiado para el examen y eso sea lo que afecta la calificación.

Gráfico 6: Diagrama de dispersión del número medio de consultas médicas anuales y la edad.

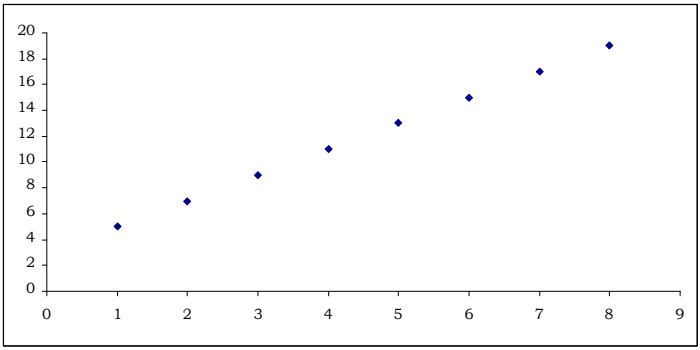


Este conjunto de puntos muestra una tendencia *no lineal*, eso no implica que las variables no estén relacionadas; por el contrario, la relación existe, pero no es lineal. Estos puntos, en lugar de ser aproximados por una línea recta, lo serían con una curva con forma de parábola. No nos ocuparemos aquí de relaciones no lineales. Limitaremos nuestro análisis a relaciones lineales, debido a que es muy frecuente usarlas como primera aproximación a la forma que tiene la relación entre dos variables y porque a menudo, cuando se trabaja con relaciones no lineales, es posible realizar transformaciones de las variables para lograr relaciones lineales.

Para analizar la intensidad de la relación lineal entre dos variables (ambas medidas a nivel intervalar o proporcional) calcularemos un coeficiente comparable a los que hemos visto hasta aquí, que tendrá una interpretación similar a la del coeficiente de correlación por rangos de Spearman. Este coeficiente se llama **coeficiente de correlación r de Pearson** y es uno de los de mayor utilización cuando las variables que se analizan alcanzan el nivel de medición que autoriza su cálculo. Este coeficiente va a medir qué tan bien se puede aproximar el conjunto de puntos con una función lineal y va a depender de lo que antes llamamos el "achatamiento" de la elipse. Será grande (próximo a 1 ó a -1) si las variables están muy relacionadas linealmente, es decir, si la nube de puntos se elonga hacia una línea; y será pequeño (próximo a cero) si las variables guardan poca relación lineal, es decir si la nube de puntos tiene forma redondeada. Será positivo y elevado (próximo a 1) si valores pequeños de una variable están acompañados de valores pequeños de la otra y valores grandes de una siguen a valores grandes de la otra, como sucedió en el ejemplo del gráfico 3.

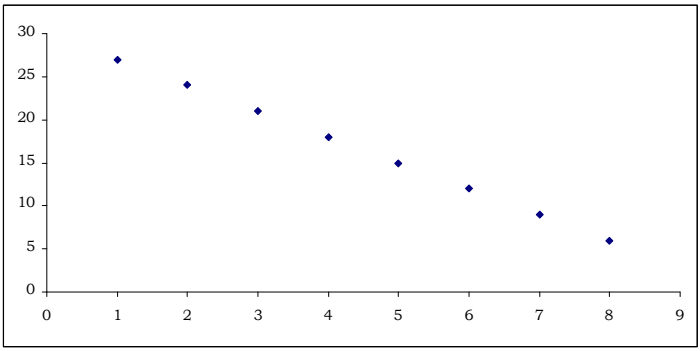
Será negativo y elevado (próximo a -1) si los valores grandes de una de las variables acompañan a los pequeños de la otra y viceversa, como en el gráfico 5. La correlación será perfecta positiva (r=1) si todos los puntos se ubican sobre una recta creciente:

Gráfico 7: Ejemplo de situación ideal con todas las observaciones alineadas en una recta creciente, por lo que r=1



Y será perfecta negativa (r=-1) si todos los puntos se ubican sobre una recta decreciente:

Gráfico 8: Ejemplo de situación ideal con todas las observaciones alineadas en una recta decreciente, por lo que r=-1



Ambas son situaciones ideales que no se encuentran en la realidad, constituyen el límite de la intensidad que pueden alcanzar las relaciones directas o inversas.

Las unidades en que se miden las variables que se relacionan pueden ser muy diferentes, en el ejemplo de ansiedad y

resultado de los exámenes, la primera se medía en una escala de 0 a 100 y la segunda de 0 a 10, por lo que un valor elevado de la primera sería 95 y uno elevado de la segunda, 9. Esto impide que se comparen directamente los valores grandes con los grandes y los pequeños con los pequeños. Vamos a usar un recurso que presentamos al final del capítulo 3, las puntuaciones *z*, que nos indican a cuantas desviaciones estándar se encuentra cada observación de la media, es decir que nos indica si se trata de un valor grande (muy superior a la media) o pequeño (muy inferior a la media) o intermedio (semejante a la media), sin tener unidades, por lo que permite la comparación de elementos que pueden tener cualquier unidad de medida[38]. Recordemos que para los valores bajos de la variable (menores a la media), el puntaje *z* es negativo y es positivo para los valores altos (superiores a la media). Si dos variables están correlacionadas positivamente (altos con altos y bajos con bajos), entonces sus puntajes *z* se corresponderán positivos con positivos y negativos con negativos. Si para cada sujeto multiplicamos los puntajes *z* de las dos variables que se relacionan, obtendremos siempre un resultado positivo, ya sea porque multiplicamos dos números positivos ($+\times+=+$) o dos negativos ($-\times-=+$). Si luego sumamos esos productos para todos los sujetos obtendremos un número alto positivo.

A la inversa, si dos variables se correlacionan negativamente los productos de sus puntajes *z* serán negativos, porque los valores altos de una irán con los bajos de la otra (que equivale a positivos con negativos, y $+\times-=-$) y bajos con altos (que es lo mismo que negativos con positivos y $-\times+=-$). Cuando sumemos estos productos para todos los casos tendremos un número alto y negativo.

Si las variables no estuvieran correlacionadas, habría casos en el que un valor alto de una variable se acompaña de uno alto de la otra y casos en que un valor alto va seguido de uno bajo, algunos productos de *z* serían positivos y otros negativos y entonces, al sumarlos, obtendríamos un número bajo, que puede ser positivo o negativo.

Entonces el producto de las puntuaciones *z* ofrece un resultado que será alto y positivo si las variables tienen una correlación alta y directa, dará alto y negativo si la correlación es inversa y dará un resultado pequeño (que puede ser positivo o negativo) si no están correlacionadas.

[38] En el lenguaje de Aron y Aron (2001), el puntaje *z* es una buena manera de comparar manzanas con naranjas.

Haciendo uso de este razonamiento, el coeficiente de correlación de Pearson se calcula como[39]:

$$r = \frac{\sum_{i=1}^{n} z_{x_i} * z_{y_i}}{n-1}$$

Donde z representan los desvíos estándar de las variables x e y, n es el total de observaciones y los subíndices i corresponden a cada una de ellas. El signo de suma señala que ésta debe extenderse desde el primer caso ($i=1$) hasta el último (n).
Como en el caso del coeficiente de Spearman, el campo de variación del coeficiente de Pearson es el intervalo -1, 1.
Para el cálculo del coeficiente de Pearson, no es necesario que las dos variables tengan las mismas unidades, porque se usan los puntajes z, que carecen de unidades. No hay inconveniente en correlacionar el peso (en kilogramos) con la talla (medida en centímetros).

A continuación, se presenta un ejemplo de cómo calcular el coeficiente de correlación de Pearson para evaluar la relación entre dos variables. Las variables seleccionadas para el ejemplo son: 1) puntaje obtenido en una escala de inteligencia lógico-matemática y 2) cantidad de ejercicios correctamente realizados en una prueba de matemática. El fragmento de la matriz de datos correspondiente es el siguiente:

[39] Esta expresión puede encontrarse un poco diferente en algunos manuales. Si las desviaciones estándar se calcularan con denominador n (como si correspondieran a observaciones provenientes de toda la población), entonces la fórmula de r llevaría denominador n también. Aquí mantenemos el modo de cálculo de la desviación estándar muestral con denominador $n-1$ y por eso esta fórmula lo lleva así también.

Sujetos	Puntaje escala inteligencia lógico-matemática	Ejercicios correctamente realizados
	X	Y
1	46	7
2	44	2
3	56	7
4	57	8
5	30	2
6	60	9
7	45	5
8	43	1
9	64	9
10	32	3

Después de obtener la media y desviación estándar de las medidas del puntaje en la escala de inteligencia lógico-matemática ($\bar{x} = 47{,}7\ s_x = 11{,}44$) y de cantidad de ejercicios matemáticos correctamente realizados ($\bar{y} = 5{,}3\ s_y = 3{,}09$), convertimos las observaciones brutas en puntuaciones z. Para ello se debe calcular la diferencia entre la puntuación bruta original y la media del grupo, y al resultado de esta operación dividirlo por la desviación estándar del grupo. La transformación a puntaje z se obtiene como vimos en el capítulo 3:
Para el puntaje en inteligencia lógico – matemática (x):

$$z_x = \frac{x - \bar{x}}{s_x}$$

Para la cantidad de ejercicios correctamente realizados (y):

$$z_y = \frac{y - \bar{y}}{s_y}$$

Entonces, el puntaje z del sujeto 1 para cada escala se obtiene de la siguiente manera:

$$z_{x_1} = \frac{46 - 47{,}7}{11{,}44} = -0{,}15$$

$$z_{y_1} = \frac{7 - 5{,}3}{3{,}09} = 0{,}55$$

Y del mismo modo para cada uno de los sujetos observados, para obtener la siguiente tabla:

Sujetos	Puntaje escala inteligencia lógico-matemática	Ejercicios correctamente realizados	$z_x * z_y$
	z_x	z_y	
1	-0,15	0,55	-0,08
2	-0,32	-1,07	0,35
3	0,73	0,55	0,40
4	0,81	0,87	0,71
5	-1,55	-1,07	1,65
6	1,08	1,20	1,29
7	-0,24	-0,10	0,02
8	-0,41	-1,39	0,57
9	1,42	1,20	1,70
10	-1,37	-0,74	1,02

El numerador del coeficiente r de Pearson se obtiene sumando la última columna:

$$\sum z_x * z_y = 7{,}63$$

y solo queda dividir este número por *n-1=10-1=9* y obtenemos $r = 0{,}85$.
El coeficiente dio positivo, por lo que la relación es directa. Como se habría esperado: las personas con mayor puntaje en la escala de inteligencia lógico–matemática, son quienes tienen una mayor cantidad de ejercicios correctamente realizados.
Además, el valor 0,85 es elevado, lo que indica que la relación entre las dos variables es intensa.

Cuando esta operación se solicita a InfoStat®, se obtiene:
Coeficientes de correlación

Correlacion de Pearson: coeficientes\probabilidades

	int. logico matem	ejercicios bien
int. logico matem	1,000	0,002
ejercicios bien	**0,847**	1,000

De esta salida nos interesa el valor del coeficiente (destacado en negrita), que aparece en la primera intersección de las dos variables que correlacionamos. Como la salida ofrece tres decimales, el resultado difiere levemente del que obtuvimos manualmente, si se redondea a dos decimales es 0,85. El formato de la salida es el mismo que cuando se obtiene el

coeficiente de Spearman, por ahora solo leemos el valor que señalamos en negrita.

La decisión acerca de considerar como grande o pequeño al valor de un coeficiente de asociación o de correlación depende del tipo de variable con que se esté trabajando y en especial de la forma en que son medidas esas variables. En buena medida, el uso de estos coeficientes es comparativo y puede ser muy valioso saber si una variable se asocia (o se correlaciona) más con una que con otra. Cuando tratamos de explicar un fenómeno y formulamos hipótesis sobre varios factores, es útil saber cuáles de ellos se asocian más intensamente con ese fenómeno.
Se han establecido algunos valores de referencia, según los cuales la correlación se considera nula si $r < 0{,}10$, pequeña si $0{,}10 \leq r < 0{,}30$, media si $0{,}30 \leq r \leq 0{,}50$ y grande si $r > 0{,}50$. Para Cohen (1988), estos criterios son un tanto arbitrarios y siempre debe considerarse al coeficiente de correlación en contexto.

En el análisis de la relación lineal entre dos variables, el coeficiente de Pearson puede ofrecernos una interpretación más detallada de la incidencia de una variable sobre la otra. En efecto, cuando este coeficiente se eleva al cuadrado, se obtiene un número que se llama **coeficiente general de determinación**, que se indica como R^2 y que mide la parte de la varianza que es compartida por las dos variables. Esta interpretación puede no ser clara en este momento, y volveremos sobre ella hacia el final del capítulo.
Cuando la relación es asimétrica y una variable opera como antecedente y la otra como consecuente, el coeficiente general de determinación mide la parte de la varianza de la variable consecuente que se explica por la antecedente. O bien la parte de la variabilidad de la variable dependiente que puede atribuirse a la variable independiente. Así, en nuestro ejemplo: $R^2 = 0{,}85^2 = 0{,}72$, lo podemos indicar como 72% y quiere decir que el 72% de la variabilidad total que aparece en el número de ejercicios correctamente realizados, se explica por el puntaje alcanzado en la escala de inteligencia lógico matemática. Así, con este coeficiente identificamos la magnitud del aporte que una variable hace a explicar los cambios de la otra. Como hemos señalado ya, los hechos que observamos obedecen a una multiplicidad de "causas" o factores explicativos, por esa razón es muy valioso disponer de un coeficiente como R^2, que nos indica qué parte de los cambios de lo observado pueden atribuirse a un determinado factor explicativo.

En el apartado siguiente volveremos sobre el coeficiente general de determinación en una aplicación más general.

Cuadro resumen de los coeficientes mencionados

Nivel de medición de las variables	Coeficiente	Rango de variación	Lectura
Nominal dicotómicas ambas	Q de Kendall	Desde -1 hasta 1	No importa el signo, la relación es fuerte si es cercano a 1 ó a -1 y débil si está cerca de 0
Nominal	C de Pearson	Desde 0 hasta C_{max}, que depende de la dimensión de la tabla	Más intensa si es próximo a C_{max}
	V de Cramer	Desde 0 hasta 1	Más intensa si es próximo a 1
Ordinal	r_s de Spearman	Desde -1 hasta 1	El signo indica la dirección, positivo es directa, negativo es inversa. Fuerte si es cercano a 1 ó a -1 y débil si está cerca de 0
Proporcional	r de Pearson		

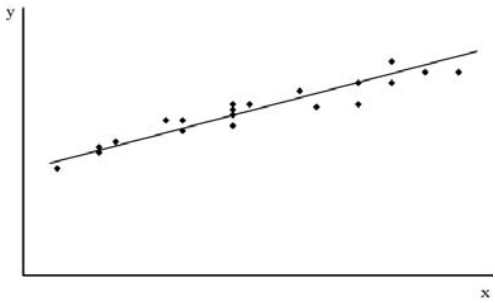
La forma de la relación

En este capítulo hemos trabajado con los coeficientes que miden la intensidad de la relación entre dos variables pero no hemos hecho referencia a la simetría o asimetría de la relación. No nos preocupamos por distinguir si una variable explica a otra o si se trata de variación conjunta, salvo la breve referencia a R^2 al final del apartado anterior. Ahora nos concentraremos en relaciones asimétricas: aquellas en las que es posible identificar a una de las variables como antecedente y a la otra como consecuente (o como independiente y dependiente, en el contexto del diseño experimental). Se trata de relaciones que se dirigen a explicar una variable (la consecuente) a partir de los valores de la otra (la antecedente). Por ejemplo, cuando preguntamos si una droga es efectiva para tratar la depresión, buscamos la relación entre las diferentes dosis de la droga y la reducción de síntomas de la depresión, por ejemplo a través del puntaje alcanzado en una prueba que la evalúa. O también, si preguntamos por el efecto del nivel de ansiedad (variable antecedente) sobre los resultados que se obtienen en un examen (variable consecuente, a explicar).

Cuando las variables tienen nivel de medición proporcional, es posible representar la relación con un diagrama de dispersión y, como hemos visto, cuanto más intensa es la relación (coeficiente de correlación de Pearson cercano a 1), tanto más se aplana la nube de puntos, yendo hacia una tendencia lineal, aproximándose a una alineación a lo largo de una recta.

En este apartado avanzamos un paso más en el análisis de la relación entre variables: cuando los puntos del diagrama de dispersión tengan una disposición semejante a una recta (creciente, como en el gráfico 3 o decreciente, como en el 5), podremos buscar la función lineal que mejor aproxima esos puntos. Usaremos a partir de aquí una notación general: llamaremos x a la variable antecedente (o independiente) e y a la consecuente (o dependiente). Porque la relación se supone asimétrica, esperamos que x tenga efectos sobre y (la dosis de droga sobre la depresión, o la ansiedad sobre los resultados del examen, en los ejemplos mencionados). Se trata de modelar la nube de puntos a través de una recta, como en el gráfico siguiente:

Gráfico 9: Ejemplo de función lineal que aproxima los puntos con una tendencia creciente



Aquí, la recta que aproxima los puntos está trazada de modo que equilibre lo que los puntos se apartan de ella por encima y por debajo. La búsqueda de esa función lineal implica proponer un **modelo** para la forma de la relación entre las dos variables, veremos qué tan realista resulta suponer que las dos variables se relacionan de manera lineal. Antes de eso será necesaria una breve referencia a esta función.

La **función lineal** tiene una expresión matemática como la siguiente $y = b_0 + b_1 * x$ en la que x e y son las variables cuya relación analizamos y los números b_0 y b_1 son valores fijos que determinan cuál es la recta de la que hablamos. Hallar la recta implica encontrar esos dos números b_0 y b_1. Una vez que están determinados, se conoce la recta y se la puede trazar.

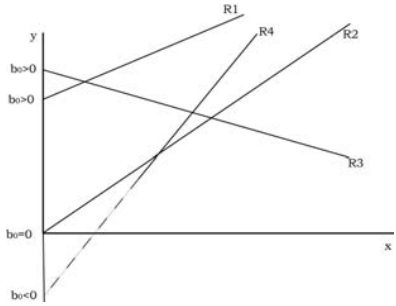
La ordenada al origen de la recta

El número b_o se llama ordenada al origen y representa el valor de y cuando x vale cero. Eso puede verse fácilmente en la expresión de la recta cuando se reemplaza a x por cero, así se obtiene:

$$y = b_0 + b_1 * x = b_0 + b_1 * 0 = b_0$$

Entonces, si $x = 0$, tendremos $y = b_0$. Como los ejes tienen el cero en el punto en que se cortan, gráficamente esta ordenada al origen se ubica sobre el eje y (por ser una ordenada), como indican las diferentes rectas en el gráfico siguiente.

Gráfico 10: Ejemplos de funciones lineales con diferentes valores de la ordenada al origen

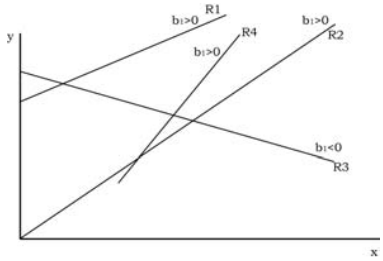


Las rectas R1 y R3 tienen ordenada al origen positiva ($b_0 > 0$). La ordenada al origen de R3 es mayor que la de R1, porque su b_0 está más arriba en el eje de ordenadas. A los efectos de la ordenada al origen, no hay ninguna diferencia en que R1 vaya subiendo y R3 baje.
La recta R2 pasa exactamente por el origen de coordenadas, por lo que su ordenada al origen es cero, $b_0 = 0$.
La recta R4 ha sido prolongada para llegar a cortar al eje de ordenadas y lo hace en un valor negativo ($b_0 < 0$).

Según las variables con que se trabaje, a veces b_0 no tiene interés, porque no se consideran los valores negativos o bien porque no tiene sentido que la variable antecedente sea cero ($x=0$).
En los ejemplos que hemos mencionado hay diferentes situaciones con respecto al valor de b_0. En la relación *dosis droga–depresión*, el valor cero para la dosis es la no-administración de la misma. Aquellos sujetos que tienen $x = 0$ son quienes no recibieron la droga y la ordenada al origen será el valor que hallemos en la escala de depresión (variable y) para quienes no tomaron la droga (a dosis cero). Por el contrario, no es posible considerar un valor cero de la ansiedad en el segundo ejemplo, por lo que allí no nos interesamos por la ordenada al origen.

La pendiente de la recta
El otro número que determina de qué función lineal se trata, es b_1 que se llama pendiente y gráficamente indica la inclinación de la recta: su valor es responsable de que la recta "suba" o "baje", siempre mirándola de izquierda a derecha.

Gráfico 11: Ejemplos de funciones lineales con diferentes valores de la pendiente



Las rectas R1, R2 y R4 son crecientes, van aumentando hacia la derecha (a medida que x crece), por eso la pendiente es positiva ($b_1 > 0$).
La recta R3 desciende, es una función decreciente, porque a medida que x aumenta y disminuye y la pendiente es negativa ($b_1 < 0$).
Vemos entonces que la pendiente depende de que sea una relación directa o inversa. Cuando es directa, x crece e y crece y la pendiente es positiva; cuando es inversa, x crece e y disminuye y la pendiente es negativa. Esto es lo mismo que sucede con el coeficiente de Pearson: positivo indica relación directa y negativo, inversa. Por esta razón, b_1 (la pendiente de la recta) siempre tiene el mismo signo de r, porque en ambos casos el signo indica si se trata de una relación directa o inversa.

Además del gráfico, el significado analítico de la pendiente es muy importante, porque indica en cuánto varía y por cada unidad que aumenta x. b_1 mide cuánto cambia la variable consecuente (dependiente), cuando la variable antecedente (independiente) cambia en una unidad. Como b_1 puede ser positiva o negativa, el cambio en y puede ser en dirección de aumentar cuando x aumenta o de reducirse. En el ejemplo de la relación *dosis droga-depresión*, se esperaría que el valor de b_1 fuera negativo, porque mide en cuánto se reduce la depresión (medida con el puntaje correspondiente a la escala que se use) por cada unidad que se aumente la dosis. Del mismo modo con la *ansiedad* y el *resultado del examen*, se espera que los sujetos con mayor ansiedad alcancen resultados menores en el examen, por lo que la relación se espera que sea inversa, con pendiente negativa y que la recta sea decreciente.
Por el contrario, si observamos la relación entre las *horas dedicadas al estudio* —como variable antecedente— y el *resultado del examen* —como consecuente—, esperaríamos una relación directa, una

pendiente positiva ($b_1 > 0$), una recta creciente que indica cómo aumenta el resultado del examen a medida que se dedican más horas al estudio.

La obtención de la recta de regresión

Para encontrar b_0 y b_1 y determinar así la función lineal que corresponde a nuestra recta de regresión, deben usarse los puntos del diagrama, es decir los pares ordenados correspondientes a cada caso. Para hacerlo usaremos las fórmulas que mostramos a continuación pero, como antes solo será para ver el modo de usarlas, luego lo pediremos a InfoStat®. Llamando x_i e y_i a cada valor de cada par ordenado y n al número total de observaciones, la expresión para calcular la pendiente de la recta es:

$$b_1 = \frac{n * \sum_{i=1}^{n} x_i * y_i - (\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{n * \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Una vez que conocemos la pendiente, se puede hallar la ordenada al origen haciendo:

$$b_0 = \bar{y} - b_1 * \bar{x}$$

Donde $\bar{x}$ e $\bar{y}$ son las medias de x y de y respectivamente.

Vamos a aplicar estas expresiones para encontrar la función lineal que mejor ajusta los puntos del ejemplo en el que relacionamos el *puntaje en la escala de inteligencia lógico-matemática* con el *número de ejercicios correctamente realizados*. Tratamos de manera asimétrica a esta relación y tomamos al puntaje de la escala de inteligencia lógico-matemática como antecedente (x) y al número de aciertos como consecuente (y). Es decir que en nuestro modelo estamos tratando de explicar el *número de aciertos* a partir del *puntaje en la escala de inteligencia lógico-matemática*.

Para facilitar el uso de la expresión del cálculo de la pendiente, agregamos dos columnas adicionales a los valores de las dos variables: la de los productos de cada x por cada y, y la de las x al cuadrado, del siguiente modo:

	x_i	y_i	$x_i * y_i$	x_i^2
	46	7	322	2116
	44	2	88	1936
	56	7	392	3136
	57	8	456	3249
	30	2	60	900
	60	9	540	3600
	45	5	225	2025
	43	1	43	1849
	64	9	576	4096
	32	3	96	1024
Sumas de las columnas	477	53	2798	23931

Tenemos entonces

$$\sum_{i=1}^{10} x_i = 477\,,\ \sum_{i=1}^{10} y_i = 53\,,\ \sum_{i=1}^{10} x_i * y_i = 2798\,,\ \sum_{i=1}^{10} x_i^2 = 23931$$

Reemplazando, obtenemos la pendiente de la recta:

Ahora

$$b_1 = \frac{n*\sum_{i=1}^{n} x_i * y_i - (\sum_{i=1}^{n} x_i)*(\sum_{i=1}^{n} y_i)}{n*\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{10*2798 - 477*53}{10*23931 - 477^2} = \frac{2699}{11781} = 0,23$$

Esta pendiente es positiva, como lo había sido r, y eso indica que la relación es directa. La pendiente además nos informa que por cada punto adicional en la variable antecedente (puntaje escala inteligencia lógico-matemática) se espera que se incremente en 0,23 el número de ejercicios bien resueltos.

Las medias de *x* y de *y* habían sido calculadas cuando las necesitamos para *r*: $\bar{x} = 47,7$ e $\bar{y} = 5,3$ por lo que la ordenada al origen de la recta es:

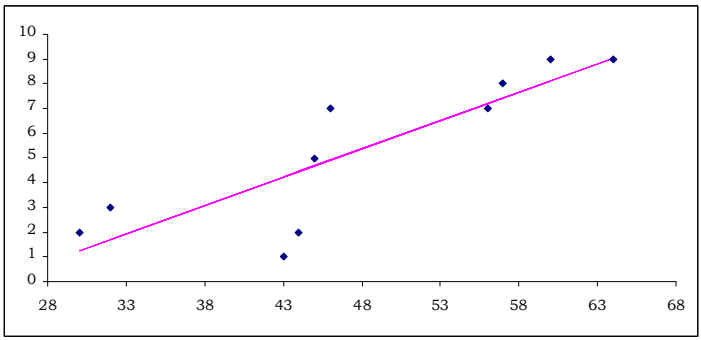
$$b_0 = \bar{y} - b_1 * \bar{x} = 5,3 - 0,23 * 47,7 = -5,63$$

El valor de la ordenada al origen no tiene interés en este ejemplo —salvo para el trazado de la recta—, porque sería el número de aciertos

esperado (que resultan negativos, es decir sin interpretación posible) para alguien con inteligencia cero, lo cual no está definido.

Conociendo la pendiente y la ordenada al origen, podemos escribir la ecuación de la recta: $\hat{y} = -5{,}63 + 0{,}23 * x$. Esta es la función lineal que describe los cambios de y a partir de los de x. Hemos escrito a y con una indicación especial, un circunflejo: $\hat{y}$, vamos a llamarla "y estimada" y es la que vamos a usar para trazar la recta:

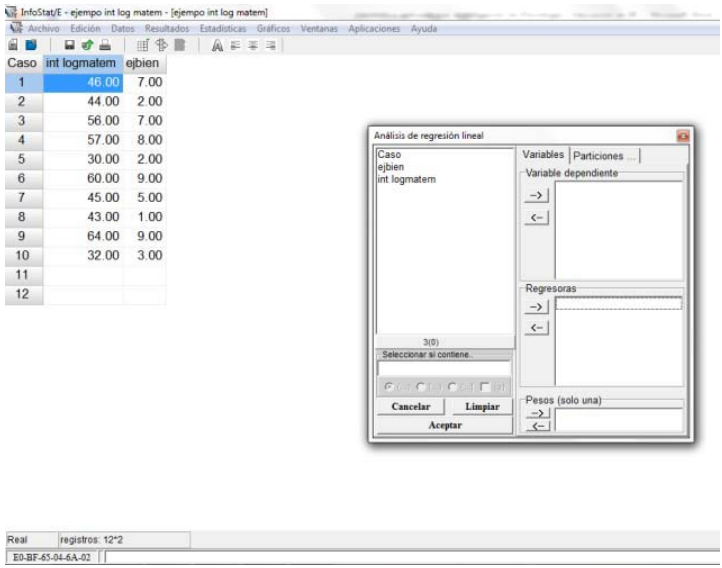
Gráfico 12: Diagrama de dispersión de la relación entre el puntaje en la escala de inteligencia lógico-matemática y el número de ejercicios de matemática correctamente realizados y la recta de regresión que mejor ajusta los puntos



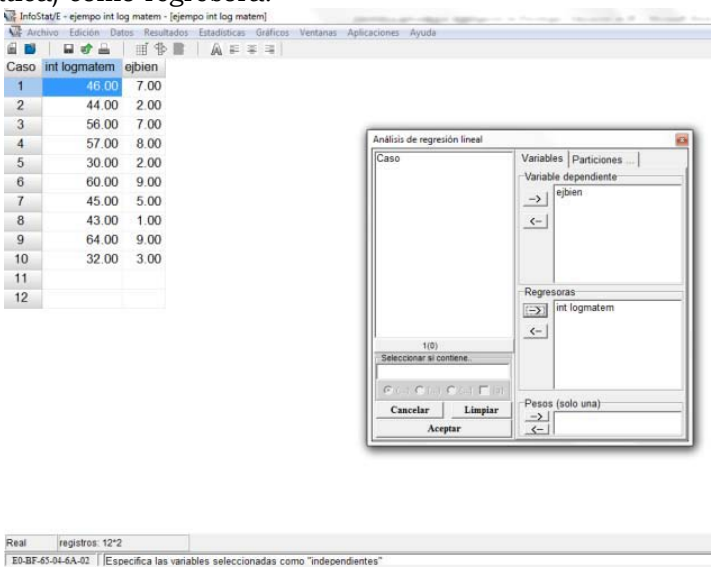
La recta que hemos encontrado usando las fórmulas de arriba es la que hace mínimos los cuadrados de las distancias de cada punto a la recta[40], por eso, a esta también se la llama **recta de mínimos cuadrados**.

Para hacer esta operación con InfoStat® hay que tener en cuenta que a las variables las llama "dependiente" a la consecuente (y) y "regresora" a la antecedente (x) y que a la ordenada al origen se la llama "constante". En el menú "estadísticas", solicitamos regresión lineal y disponemos de este cuadro para seleccionar nuestras variables:

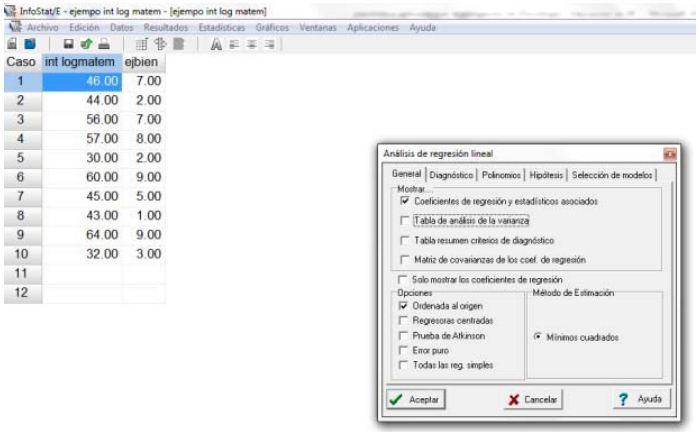
[40] No es posible poner como condición que la recta haga mínimas las distancias porque hay puntos por encima y por debajo, por lo que la suma de las distancias se hace cero (igual a lo que sucedió con la suma de los desvíos alrededor de la media y que llevó a usar sus cuadrados para definir la varianza). Por esa razón se usan los cuadrados de las distancias.



Elegimos ejbien (número de ejercicios bien resueltos) como variable dependiente y int logmatem (puntaje en la escala de inteligencia lógico matemática) como regresora.



Luego de aceptar, se nos ofrecen opciones, de las que solo pedimos "coeficientes de regresión y estadísticos asociados" y "ordenada al origen".



El formato de la salida es el siguiente:

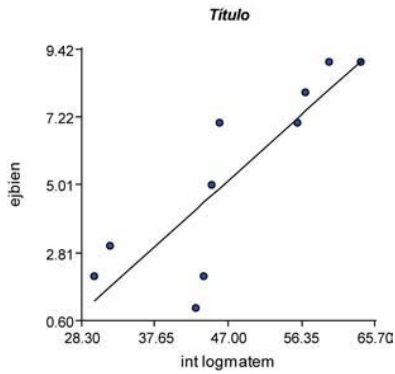
Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
ejbien	**10**	**0.72**	0.68	4.17	43.24	44.15

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	**-5.63**	2.48	-11.35	0.10	-2.27	0.0531	
int logmatem	**0.23**	0.05	0.11	0.35	4.51	0.0020	19.23

En esta salida tenemos dos partes. De la primera consideraremos por ahora: el número de observaciones *N=10*, y el coeficiente general de determinación $R^2=0,72$, al que ya habíamos calculado antes como cuadrado del coeficiente de Pearson.
De la segunda parte nos interesan los números -5,63 y 0,23. El primero está identificado como "constante", y es la ordenada al origen: b_0. El segundo, indicado como *int logmatem* (el nombre de la variable antecedente), es la pendiente de la recta: b_1.
El gráfico que produce InfoStat® es el siguiente, que permite editarse para colocar título y rótulos diferentes a los ejes.



Además de mostrarnos la forma del modelo, la recta de regresión sirve para hacer estimaciones de valores no observados, porque nos ofrece valores de $\hat{y}$ para cada x que reemplacemos. Por ejemplo, si preguntamos por la cantidad de ejercicios bien resueltos que se esperan en alguien que alcanzó 55 puntos en la escala de inteligencia lógico-matemática, respondemos reemplazando en la función el valor de $x=55$ y resulta:

$$\hat{y} = -5{,}63 + 0{,}23 * x = -5{,}63 + 0{,}23 * 55 = 7{,}02$$

Que puede redondearse a 7. Este es el valor estimado del número de aciertos para alguien con 55 puntos en la escala de inteligencia lógico matemática.
Estas estimaciones son muy valiosas para hacer predicciones sobre valores que no han sido observados, por ejemplo hacia el futuro. Ejemplos muy útiles de esta aplicación son las proyecciones de población, y más específicamente las de matrícula escolar, que ofrecen estimaciones del volumen de alumnos que se prevé para años próximos.

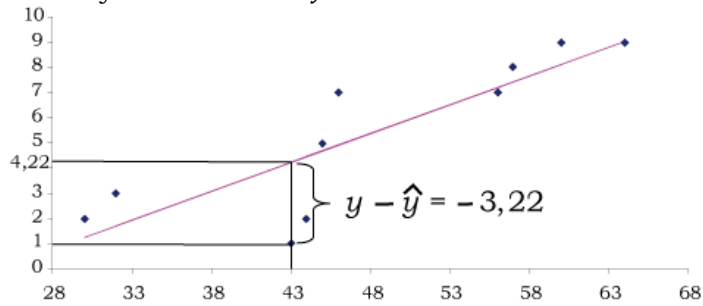
Cuando se reemplaza cada uno de los x observados en la función, se encuentran las estimaciones para cada uno de ellos. En la tabla siguiente indicamos cada uno de los pares ordenados como fueron observados y agregamos los valores de $\hat{y}$ estimados a través de la función lineal[41]; por último, restamos los valores estimados de y, de los reales, para ver las diferencias entre los que la recta estima y los que hemos observado:

[41] Para calcular los valores de y estimado ($\hat{y}$) hemos conservado más decimales en b_0 y b_1 que los mostrados.

x	y	$\hat{y}$	$y - \hat{y}$
46	7	4,91	2,09
44	2	4,45	-2,45
56	7	7,20	-0,20
57	8	7,43	0,57
30	2	1,24	0,76
60	9	8,12	0,88
45	5	4,68	0,32
43	1	4,22	-3,22
64	9	9,03	-0,03
32	3	1,70	1,30

La última columna mide la distancia que hay entre cada punto y la recta. Es un indicador de la calidad del ajuste que hace la función lineal de los puntos. Cuando esas distancias son pequeñas tenemos una recta que ajusta mejor los puntos que cuando las distancias son grandes. Como vemos, en este ejemplo hay algunas positivas, que corresponden a los puntos que están encima de la recta, y otras negativas, las de los puntos por debajo de la recta. Veamos en el gráfico siguiente la ubicación de uno de estos puntos, por ejemplo el que corresponde al par observado (43; 1), al que la recta estima con el valor $\hat{y} = 4{,}22$:

Gráfico 13: Ubicación gráfica de la diferencia entre el valor observado de y observado y su estimación $\hat{y}$



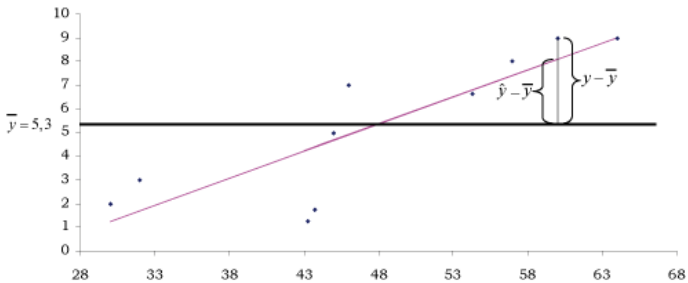
Este desvío es negativo porque el punto está debajo de la recta, la estimación $\hat{y}$ es mayor que el valor observado, y.

La suma de todos estos desvíos es cero, por haber pedido a la recta la condición de equilibrar los puntos. Estas diferencias son los errores

que se cometen al estimar a través de la función lineal. Para verlos como tales, observemos que si la correlación fuera perfecta (positiva o negativa) como en los gráficos 7 y 8 todos los puntos estarían sobre la recta y coincidirían los valores de y con los de $\hat{y}$, por lo que las diferencias de la última columna serían todas cero, no habría error en una relación perfecta. Como hemos dicho, esa es una situación ideal, que no puede observarse en la realidad: en los casos reales siempre hay apartamientos de los puntos a la recta, que constituyen el error de estimación. Para llegar a una medida de la calidad de nuestro modelo, es decir una medida de qué tan bueno es el ajuste que la recta hace sobre los puntos realmente observados, trabajaremos sobre la dispersión, a través de la varianza.

En primer lugar, la variable y tiene su varianza, que mide lo que los valores se apartan de la media, vamos a llamarla s_y^2 (la varianza de y). En segundo lugar, los $\hat{y}$ también se apartan de la media y esas distancias pueden resumirse en otra varianza, la que mide las distancias desde los $\hat{y}$ hasta la media de y, la llamaremos $s_{\hat{y}}^2$ (la varianza de $\hat{y}$). Así entonces resumimos los desvíos de la variable hasta la media, con la varianza de y, y los desvíos de las estimaciones hasta la media, con la varianza de $\hat{y}$. Si se traza una recta horizontal para ubicar a la media de y ($\bar{y}$ en 5,3) y se recuerda que los valores de y son los que están en los puntos realmente observados, mientras que los de $\hat{y}$ están sobre la recta, estos desvíos pueden verse gráficamente así:

Gráfico 14: Ubicación gráfica de los alejamientos (desvíos) de y observada y de $\hat{y}$ estimada, hasta la media de y ($\bar{y}$)



Lo que cada observación se aleja de la media es la diferencia $y - \bar{y}$. Cuando se consideran todas ellas, su medida es la varianza de y (s_y^2).

Lo que se desvía la estimación (ubicada sobre la recta) de la media es $\hat{y} - \bar{y}$, que cuando se extiende a todos los puntos se resume en la varianza de $\hat{y}$ ($s^2_{\hat{y}}$).

La calidad del ajuste se aprecia en la proximidad que la recta tiene a los puntos, en el caso ideal (si *r=1* ó *r=-1*), las distancias son todas cero. En una situación real, el ajuste será tanto mejor cuanto más cerca estén los puntos de la recta; cuanto más pequeñas sean esas distancias. Para medir esta calidad se usa el cociente entre las dos varianzas anteriores, que resulta ser el ya mencionado coeficiente general de determinación: $R^2 = \frac{s^2_{\hat{y}}}{s^2_y}$.
Al que ahora calculamos como la varianza de $\hat{y}$ sobre la varianza de y, por lo que mide qué porción de la variabilidad total de y (el denominador) representa la variabilidad de $\hat{y}$, puede leerse como la parte de toda la variabilidad de la variable dependiente que es explicada por el modelo lineal.
Este cociente no puede ser negativo ni mayor que 1, porque el numerador es menor o igual que el denominador. Solo vale 1 en el caso de una asociación perfecta, en que las varianzas son iguales —porque los puntos están sobre la recta—, *r=1* ó *r=-1* y, en consecuencia $R^2=1$.

El coeficiente general de determinación mide la proporción de los cambios de y (expresados con la varianza) que se explican a través de la función lineal. Por eso es muy valioso, porque cuantifica el peso relativo de la variable x (a través de la función lineal) en la explicación de y.
Podemos así decir que, por ejemplo, el 30% de las diferencias en el rendimiento escolar de los alumnos de primaria se explica por la educación de sus padres, o que el 60% de la disminución del puntaje en una prueba que evalúa la depresión se puede atribuir a la administración de una determinada droga. Vemos con estos ejemplos la gran potencialidad explicativa de este coeficiente.

Todo el análisis de regresión que hemos desarrollado hasta aquí puede hacerse de manera equivalente si la relación entre las variables no es lineal. En vez de obtener la ecuación de una recta, se obtendrá la de una parábola, una cúbica o cualquier otra función que aproxime adecuadamente los pares ordenados que se observan. La definición del coeficiente general de determinación que hemos dado: $R^2 = \frac{s^2_{\hat{y}}}{s^2_y}$ no

cambia, pero los $\hat{y}$ se calcularán con la función adecuada, no con la lineal que vimos.

Sobre el coeficiente de correlación de Pearson, debe recordarse que solo es adecuado para evaluar la intensidad de la relación entre dos variables si ésta es lineal. En presencia de una relación de otro tipo, como la del gráfico 6, el coeficiente de Pearson dará un valor muy bajo, pero eso no quiere decir que la relación sea débil o inexistente, sino que el modelo lineal no es adecuado para describirla. Por ello, si se encuentra un coeficiente de Pearson muy bajo, debe explorarse la existencia de una relación no lineal, esto puede hacerse fácilmente observando cómo se disponen los puntos en el diagrama de dispersión.

Cuando la relación se modela con una función lineal —y solo en estos casos—, el coeficiente general de determinación (R^2) se calcula directamente elevando al cuadrado al coeficiente de Pearson.

Actividad práctica de repaso 5

1. La siguiente tabla bivariada muestra los resultados de una recolección de datos hecha sobre una muestra de 195 personas adultas, acerca del nivel de instrucción y el hábito regular de lectura:

		Lectura regular		
		Si	No	Total
Nivel de instrucción	Primario incompleto o menos	20	50	70
	De primario completo a secundario incompleto	10	15	25
	Secundario completo o más	70	30	100
	Total	100	95	195

A partir de ella puede formularse una hipótesis sobre de la relación entre las dos variables. Para poner a prueba esa relación calculamos las frecuencias que esperaríamos encontrar en las celdas si las variables fueran independientes, las llamadas frecuencias esperadas:

		Lectura regular		
		Si	No	Total
Nivel de instrucción	Primario incompleto o menos	**36**		70
	De primario completo a secundario incompleto	**13**		25
	Secundario completo o más			100
Total		100	95	195

Las frecuencias destacadas fueron calculadas como si fueran independientes.
a. Calcule las demás frecuencias esperadas
b. Si en esta tabla el puntaje chi cuadrado vale 29,75, calcule un coeficiente de asociación para esta relación, indique su nombre.
c. ¿Qué significa el valor hallado?

2. Ahora el nivel de instrucción (con las mismas categorías de arriba) se relaciona con la cantidad de horas al día que se dedican a ver televisión.
a. ¿Cuál es el coeficiente más adecuado para medir la intensidad de la relación?

Para ese coeficiente se encuentra un valor de -0,60.
b. ¿Qué significa el signo menos?

c. Redacte una interpretación para el valor hallado.

3. En el mismo grupo de personas se plantea la relación entre la cantidad de horas al día que se dedican a la lectura y el número de errores cometidos en una prueba de ortografía.
a. Indique en un sistema de ejes coordenados la ubicación de las variables (a cuál llamaría *x* y a cuál *y*) y la forma que podría tener la nube de puntos si se supone una relación lineal.

La función lineal que mejor ajusta los puntos tiene la siguiente forma:
*y=5-0,3*x*
b. ¿Cómo se llaman los números 5 y -0,3?
c. ¿Qué significa el signo menos?
d. ¿Cuál sería el coeficiente más adecuado para evaluar la intensidad de la relación?
e. Si para ese coeficiente se obtiene el valor -0,12, ¿Cómo se lee?

Al elevar al cuadrado ese coeficiente se obtiene 0,014
f. ¿Cómo se llama ese número?
g. ¿Cuál es su interpretación en referencia a las variables cuya relación se analiza?

Capítulo 6: Bases probabilísticas para la inferencia

Eduardo Bologna

El rol de la probabilidad en Estadística

Hasta aquí hemos trabajado sobre datos que han sido observados; Pudo haberse tratado de datos de una encuesta o un relevamiento que nosotros hayamos realizado o bien que nos haya provisto alguna fuente de buena calidad: registros de una institución educativa, historias clínicas de un hospital, etc. Se trata de información realmente recopilada, que ha sido obtenida por observación a través de algún instrumento de recolección de datos. Por el contrario, en los capítulos que siguen nos ocuparemos de lo que *no* ha sido observado, haremos *inferencias* para sacar conclusiones acerca de lo que no hemos visto. Partiremos de la información que provee una muestra y con ello generalizaremos a un conjunto mayor. Como resultado de ello buscaremos dar respuesta a preguntas como las siguientes:

-Si en un grupo de alumnos de una escuela se ven más dificultades en Matemática, ¿será esta observación válida para los alumnos de otras escuelas similares?

-Si un nuevo medicamento tiene efecto en un grupo de pacientes sobre los que se experimenta, ¿bajo qué condiciones podemos saber si también tendrá efecto en otros pacientes?

-Si en una muestra de 200 personas encuestadas el 42% dice que va a votar al partido A, ¿qué porcentaje se votos puede esperar obtener el partido A cuando sean las elecciones?

-Si en un grupo de voluntarios, elegidos para participar en un experimento se descubre una relación entre las expectativas de logro y el número de errores que se cometen en una prueba, ¿es suficiente con eso para afirmar que esa relación se mantiene entre quienes no participaron del experimento?, o dicho de otra manera ¿es esa una relación general entre esas variables (expectativa de logro y cantidad de errores)?

-Si a una muestra de pacientes psicóticos se administra una droga y se encuentra que produce estabilización en los síntomas en el 90% de ellos, ¿corresponde recomendar que esa droga sea utilizada en los pacientes diagnosticados como psicóticos?

Como vemos, se trata de llevar los resultados más allá del ámbito en el que fueron obtenidos, se trata de generalizarlos. Para hacer esto será necesario usar conceptos del terreno de la probabilidad, a ese tema nos dedicamos en este capítulo.

Sobre la relación entre la probabilidad y la inferencia, Ian Hacking cita un pasaje de un relato indio[42] en el que uno de los personajes estima "el número de hojas y frutos que hay en dos grandes ramas de un árbol frondoso. Aparentemente lo hace en base a una sola rama más pequeña que observa. Hay, según afirma, 2095 frutos. [...] y cuando le preguntan ¿cómo pudo saberlo?, responde: *Yo de los dados poseo su ciencia y así en los números diestro soy*" (Hacking, 2005, p.20). De este modo se liga en este antiguo texto la capacidad para hacer una estimación de lo que no es observado con "la ciencia de los dados", desde tan temprano hay indicios de la relación entre estimación y probabilidad.

Las respuestas a las preguntas que buscan generalidad no pueden ser certeras, como en el relato mítico, sino inciertas. A diferencia de las descripciones, que se limitan a mostrar información recopilada, las inferencias solo pueden ser afirmaciones tentativas, aproximativas, probabilísticas. La diferencia entre la certeza de una descripción y la incertidumbre de una inferencia se ve con claridad al comparar las siguientes expresiones:

A. "El tiempo promedio que tardaron estas 100 personas en responder al cuestionario fue de 12 minutos"

B. "Cuando este cuestionario sea aplicado, se espera que los encuestados tarden entre 11 y 13 minutos en responderlo, con una certeza del 95%".

Hay dos diferencias en estos enunciados, que pertenecen a distintos niveles de proximidad a los datos. Una diferencia es que el primero ofrece un valor único: los 12 minutos que se obtuvieron al promediar los tiempos de los 100 sujetos que fueron observados. Por el contrario, el segundo ofrece un intervalo: entre 11 y 13 minutos.

La segunda diferencia es que el primer enunciado *afirma* ese valor, mientras que el segundo expresa que *hay una certeza del 95%*. Esto quiere decir que no estamos seguros que el tiempo que tardarán en responder vaya a estar realmente entre 11 y 13 minutos; hay una confianza del 95% que sea así, pero no una certeza plena. Por eso,

[42] Se trata de la epopeya Mahabarata, cuya versión actual habría sido concluida hacia el año 400d.c.

puede ser que el tiempo sea, o bien menor que 11 minutos, o bien mayor que 13; y esto puede suceder con una probabilidad de 5%. Los enunciados del segundo tipo transmiten cierto grado de incertidumbre porque se refieren a casos que no han sido observados, sino inferidos. Esta incertidumbre, para la que disponemos de procedimientos que permiten cuantificarla de manera probabilística, es inherente a todo proceso de inducción, donde se requiere formular generalizaciones.

Los capítulos que vienen a continuación son los que pertenecen a la parte que llamamos *Estadística Inferencial*, como etapa posterior a lo que hemos visto hasta aquí, que llamamos *Estadística Descriptiva*. Para llegar a ello será necesario manejar algunos conceptos de probabilidad, y ése es el tema del que nos ocuparemos en este capítulo, que será de articulación entre lo descriptivo y lo inferencial.
Al hacer inferencias necesitamos de la probabilidad porque trabajamos con situaciones inciertas, que no conocemos y que tenemos dificultad para prever. Por eso empezaremos haciendo la distinción entre las preguntas que podemos responder con certeza y las que no. De las primeras, son ejemplos: ¿cuándo será el próximo eclipse de sol? ¿cómo cambia la conducta de una persona si consume una sustancia alucinógena? Sobre estas preguntas tenemos, o bien un conocimiento profundo sobre el movimiento de los astros, o bien una gran cantidad de observaciones, que nos permiten dar una respuesta certera.
Por el contrario, si preguntamos ¿cuál es el efecto sobre la personalidad, de haber tenido figuras parentales autoritarias en la niñez? ¿qué determina que algunos alumnos tengan éxito en la escuela y otros no?, solo podemos ofrecer respuestas parciales, tentativas, aproximadas. Se trata de hechos que dependen de muchos factores a los que no conocemos en su totalidad, por lo que el resultado es variable: algunas personas criadas en ambientes autoritarios desarrollan una personalidad autoritaria, otras no. En algunos alumnos, el hecho que sus padres tengan estudios elevados los ayuda a tener éxito en la escuela, pero también hay hijos de personas muy educadas que fracasan en la escuela. Hay otras razones que no está a nuestro alcance conocer en su totalidad, que inciden en la personalidad o en el resultado de la escuela. En estas situaciones, cuando no tenemos toda la información que hace falta para predecir el resultado, recurriremos a la probabilidad.
Ingresaremos al tema desde situaciones muy sencillas, desde el muy usado ejemplo de arrojar una moneda. Pero detengámonos un momento en él: si tuviéramos toda la información necesaria para predecir la trayectoria de la moneda en el aire (distancia desde donde

se arroja, fuerza que se le aplica, parte de la moneda donde se aplica esa fuerza, eventuales corrientes de aire que puedan incidir en el desplazamiento de la moneda, etc.), podríamos predecir con certeza el resultado. Esa información no está disponible, el lado del que caiga la moneda está determinado por una multiplicidad de factores, por esa razón no podemos anticipar el resultado de la tirada. A esa ignorancia la resumimos diciendo que el resultado de la tirada de la moneda "depende del azar" y llamamos al experimento de tirar una moneda "experimento aleatorio".

Es un paso muy largo ir desde este ejemplo a decir que el modo en que se desarrolle la personalidad de alguien que ha sido criado en una familia autoritaria depende del azar. Sabemos que no depende del azar, depende de muchos factores que ignoramos, por eso usaremos probabilidades en nuestra disciplina. Podremos decir que un alumno cuyos padres valoran la educación tiene una probabilidad mayor de tener éxito en la escuela, pero no podremos asegurar que lo tendrá.

Los eventos que no son azarosos no tienen que ver con probabilidades: no asignamos probabilidad a un eclipse, hay conocimiento suficiente como para saber cuándo ocurrirá. Asignamos probabilidades a hechos de cuya ocurrencia no estamos seguros. Con la probabilidad cuantificamos nuestras expectativas sobre el fenómeno. Intuitivamente, cuando decimos que algo tiene "mucha probabilidad de suceder" es porque estamos bastante seguros que sucederá.

Formas para asignar probabilidades

Asignación a priori

Podemos partir de esa idea intuitiva de probabilidad, ligada a procesos cuya ocurrencia no nos es conocida con certeza. Para evocar esta idea, el ejemplo que más a menudo se cita es el del lanzamiento de una moneda, ¿cuál es la probabilidad de obtener "cara" al arrojar una moneda? Si la respuesta es ½, debe tenerse en cuenta que eso solo será cierto si la moneda está equilibrada, es decir si tiene iguales chances de salir de un lado que del otro. Si esto es cierto, efectivamente la probabilidad de obtener cara es ½ (ó 0,50). Con idéntica condición, la probabilidad de obtener un 5 al arrojar un dado es 1/6 (ó 0,17). Esta asignación de probabilidad a los resultados de un experimento es previa a su realización, no es necesario tirar realmente la moneda: es suficiente con que tengamos razones para *suponer* que está equilibrada, para que afirmemos que la probabilidad de cara es ½. Diremos en este caso que asignamos la probabilidad *a priori*, es decir, antes de hacer el experimento.

De mismo modo sucede si el evento que nos interesa en un poco más complejo. Por ejemplo: ¿Cuál es la probabilidad de obtener un número mayor a cuatro si se tira un dado? Debido a que hay dos números mayores a cuatro (5 y 6), el evento tiene dos casos a su favor y hay seis resultados posibles, por lo que la probabilidad será: 2/6 (ó 1/3, si se simplifica la fracción).
La expresión formal de esta asignación de probabilidades es

$$P(A) = \frac{\#A}{\#\Omega}$$

En la que #A (que se lee "numeral de A") indica el número de maneras en que puede suceder el evento A, y #Ω (numeral de omega) es el número total de resultados que se pueden obtener al realizar el experimento. Ω es el conjunto de resultados posibles, es llamado *espacio muestral*. En el caso del ejemplo, el experimento es el de tirar el dado y buscar un número mayor que cuatro, #A es 2 porque son las formas en que puede obtenerse un número mayor que cuatro, y #Ω es 6, que es el número total de resultados posibles al tirar un dado.
Con este mismo razonamiento, la probabilidad de obtener un número par es 3/6 (1/2 después de simplificar), porque hay tres números pares (2, 4 y 6) en un dado.

Vamos a un caso más complejo: tiremos ahora dos dados y tomemos en cuenta la suma de los dos puntajes, a esa suma la llamaremos S. El mínimo número que puede resultar es dos (que ambos dados salgan uno) y el máximo es doce (ambos seis), entonces hay once resultados posibles de esta variable (que son: S = 2, S = 3, S = 4, S = 5, S = 6, S = 7, S = 8, S = 9, S = 10, S = 11 y S = 12), algunos de los cuales pueden suceder de varias formas. Estos resultados posibles y sus formas de obtención se ven de manera esquemática a continuación:

Esquema 1: Resultados posibles de la suma de los puntajes de dos dados

		Primer dado					
		1	**2**	**3**	**4**	**5**	**6**
Segundo dado	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Si bien los resultados posibles son 11, las formas en que estos pueden darse son 36; cada una de esas formas es un evento. El evento primer dado 5 y segundo dado 2 es diferente del evento primer dado 2 y segundo dado 5, aunque ambos conducen al mismo resultado: S=7. Más precisamente, si indicamos los eventos con pares ordenados, los eventos (1,6); (6,1); (2,5); (5,2): (3,4); (4,3) son diferentes pero todos corresponden a S=7.
Son entonces 36 los resultados posibles del experimento, por lo que #Ω = 36. Ahora podemos calcular probabilidades para diferentes resultados.
¿Cuál es la probabilidad que la suma sea 12?, lo que puede expresarse como: ¿cuál es P(S = 12)? Como esta suma solo puede lograrse si ambos dados salen 6, hay una sola manera en que se produzca el evento que nos interesa (suma doce), por lo que el #A es 1 y la probabilidad es entonces 1/36.

#A = 1, P(S = 12)=1/36

Segundo dado \ Primer dado	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

En cambio, si la pregunta es por la probabilidad de obtener un tres, hay más de una manera de llegar a ese resultado (que S = 3). La suma 3 puede resultar de 2+1 ó de 1+2, es decir que, o bien el primer dado sale 2 y el segundo 1 ó bien el primero sale 1 y el segundo 2. Hay así dos formas posibles para el evento S = 3, y #A es 2, por lo que la probabilidad es P(S = 3) = 2/36.

#A = 2, P(S = 3)=2/36

Segundo dado \ Primer dado	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Otro ejemplo, sea P(S=7). La suma de siete puede obtenerse de muchas formas: 1+6, 2+5, 3+4, 4+3, 5+2 ó 6+1. Hay seis combinaciones que conducen a S=7, en consecuencia, la probabilidad es 6/36.

#A = 6, P(S = 7)=6/36

Segundo dado \ Primer dado	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Dos casos particulares
1. Jamás se obtendrá uno (1) al sumar los resultados de dos dados, por lo que el evento S = 1 es imposible y si preguntáramos por P(S = 1), la respuesta es cero, como lo es también si pedimos P(S>12). De manera general, diremos que la probabilidad de un evento imposible es cero, decir que un evento tiene probabilidad igual a cero, equivale a decir que no puede suceder.
2. Inversamente, al tirar dos dados siempre se obtendrá un número menor que 13, por lo tanto, el evento S<13 siempre sucede. Esto es así porque S<13 incluye todos los resultados posibles, lo mismo sería si se pidiera que S<14 o que S<75, o también S>1, cualesquiera de ellas coincide con Ω. Para indicar este tipo de evento, sobre el que hay certeza absoluta de su aparición, decimos que su probabilidad es uno. Así, P(Ω)= 1. Decir que un evento tiene probabilidad uno equivale a decir que hay completa seguridad que va a suceder.

Asignación a posteriori
Consideremos ahora una situación más cercana a la experiencia: supongamos que de un curso se selecciona un alumno al azar ¿Cuál es la probabilidad que sea mujer? Aquí el supuesto de equilibrio no es válido a priori, en gran medida depende de la carrera de que se trata, las hay con muchas mujeres y con pocas. Por lo tanto no podemos suponer que es igualmente probable que resulte un varón o una mujer y no podemos asignar probabilidad ½ a cada resultado.
En otro ejemplo, si al alumno aleatoriamente seleccionado se le pregunta por el tipo de colegio del que egresó, con los resultados

posibles: “público”, “privado laico”, “privado religioso”, el resultado que se obtenga también depende del azar, porque así fue elegido el alumno, la respuesta será una u otra según cuál sea el alumno elegido y esto depende del azar. Sin embargo, no podemos asignar una probabilidad igual a cada resultado, no es lícito decir que sea 1/3, ya que puede haber más estudiantes que provengan de colegios públicos que de privados y, en consecuencia, que sea más probable encontrar alumnos provenientes de esos colegios que de los otros. En este caso no podemos asignar de antemano probabilidades a los diferentes resultados, porque no tenemos suficientes razones para suponer *la forma en que se distribuyen* las probabilidades. Si supiéramos que hay el doble de alumnos que vienen de colegios públicos, podríamos decir que el alumno elegido al azar tiene el doble de probabilidad de provenir de un colegio público que de otro tipo.
Encontramos así una relación entre la frecuencia y la probabilidad: si conocemos la distribución de frecuencias, tenemos razones para usarlas para asignar probabilidades a los resultados del experimento. Así, si una muestra de alumnos ofrece la siguiente distribución de frecuencias para el tipo de colegio del que provienen.

Tabla 1: Distribución de frecuencias del tipo de colegio del que provienen alumnos que ingresan a la universidad

	f	f'
Público	200	0,66
Privado laico	20	0,07
Privado religioso	80	0,27
Total	300	1,00

Fuente: datos ficticios para ejemplificación

Estamos autorizados a decir que la probabilidad que el alumno elegido al azar provenga de un colegio público es 0,66, que es su frecuencia relativa. Esto será válido en la medida que el número de casos sea elevado; no es posible transformar una frecuencia relativa en probabilidad si tenemos muy pocas observaciones. Más adelante volveremos sobre esta importante limitación.

Cuando atribuimos probabilidades de este modo se trata de probabilidades *a posteriori*, es decir con posterioridad a haber hecho la experiencia, luego de la observación de los resultados reales obtenidos. También se llama a estas probabilidades *empíricas*, para destacar que provienen de la experiencia.
¿Cuál es el significado del 1,00 que corresponde al total? Sin dudas, como frecuencia significa el 100% de los casos, pero como

probabilidad indica que el conjunto completo de alternativas (público, privado laico, privado religioso) tiene probabilidad 1,00; o bien que es un *evento seguro*. La expresión coloquial para este valor es que cuando se selecciona un alumno al azar, éste debe provenir de alguno de los tres tipos de colegio indicados, entonces el valor uno responde a la pregunta "¿cuál es la probabilidad de encontrar un alumno que provenga de un colegio público, privado laico o privado religioso?", la respuesta es que es seguro que de alguno de esos tipos de colegio provendrá el alumno, no hay otras alternativas, es decir que las categorías son exhaustivas, como se exigió en el capítulo 1. En este caso,

$$\Omega = \{público, privado\ laico, privado\ religioso\}$$

Como vimos más arriba, el evento seguro es el que tiene probabilidad 1, por lo que la expresión formal de este enunciado es:

$$P(\Omega) = 1$$

Donde Ω indica el conjunto de todos los resultados posibles de un experimento aleatorio, o el conjunto de todas las categorías de un variable, al que hemos llamado *espacio muestral*.

La relación entre asignación a priori y a posteriori

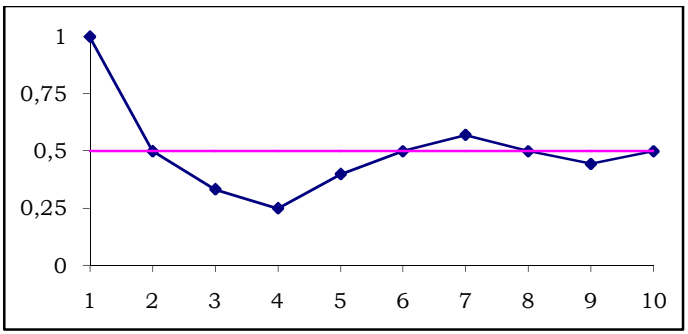
Volvamos al experimento simple de arrojar la moneda: si ésta se encuentra equilibrada será entonces correcto asignar probabilidad ½ (ó 0,50) a cada lado, lo que indica que esperamos que *a la larga* la moneda caiga la mitad de las veces cara y la mitad cruz. Destaquemos la expresión "a la larga", que quiere decir "si se arroja muchas veces". Hagamos el experimento realmente, busque usted una moneda y arrójela, digamos 10 veces. Yo lo hice y obtuve la siguiente secuencia de resultados: CXXXCCCXXC.

Elijamos nuestro lado favorito, que sea "cara" (C) y calculemos la frecuencia relativa de ese lado en cada tirada, presentamos los resultados resumidamente en la tabla 2. Cuando solo la he tirado una vez y como el primer resultado fue C, la frecuencia es 1 (una cara de un total de un resultado). A la segunda tirada, que es X, se obtiene 0,50 (una cara de dos tiradas). A la tercera, que otra vez sale X, la frecuencia de cara es 1/3 (una cara de tres tiradas) y así sigue:

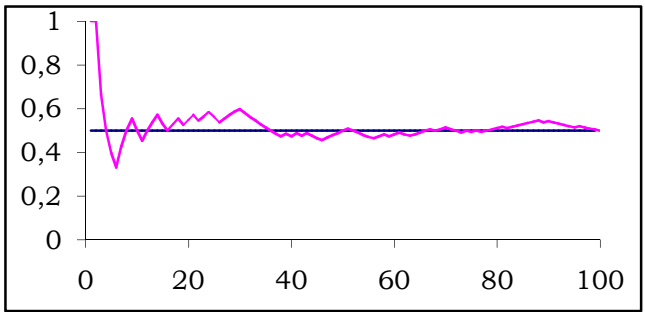
Tabla 2: Frecuencias relativas correspondientes a una secuencia de diez lanzamientos de una moneda

Tirada	Cantidad de tiradas	Resultado	Cantidad de caras acumuladas	Frecuencia relativa del evento "cara"
Primera	1	C	1	1/1 = 1,00
Segunda	2	X	1	1/2 = 0,50
Tercera	3	X	1	1/3 = 0,33
Cuarta	4	X	1	1/4 = 0,25
Quinta	5	C	2	2/5 = 0,40
Sexta	6	C	3	3/6 = 0,50
Séptima	7	C	4	4/7 = 0,57
Octava	8	X	4	4/8 = 0,50
Novena	9	X	4	4/9 = 0,44
Décima	10	C	5	5/10 = 0,50

En la distribución vemos que la frecuencia relativa de C toma valores alrededor de 0,50; al principio más lejos y luego "se va acercando" a ese número cuando hemos hecho más tiradas. La representación gráfica de este proceso es la siguiente:



Donde hemos agregado una línea que muestra el valor 0,50 que es el que habíamos calculado antes de hacer el experimento. Se aprecia que la sucesión de frecuencias relativas es tal que éstas se van acercando a la probabilidad predicha. Si la moneda se hubiese tirado una cantidad mayor de veces (cien veces, por ejemplo), el gráfico tendría una forma como la siguiente:



Resulta entonces que, si se cumple que la moneda está equilibrada, entonces las frecuencias relativas de un resultado se irán acercando a la probabilidad asignada. Dicho de otra manera, la probabilidad a posteriori converge a la probabilidad a priori. Mucha atención a esto: solo en el caso que el supuesto inicial se cumpla, es decir que la moneda se comporte como esperamos (cayendo parejo de un lado y del otro).

Si por el contrario, la moneda estuviese desequilibrada —y que fuera más frecuente el lado X que el lado C— podríamos obtener, en 500 tiradas, por ejemplo, 350 veces X y la probabilidad a posteriori será entonces P(C) = 350/500 = 0,70 y no 0,50 como sería si estuviera equilibrada. En ese caso el supuesto inicial de moneda equilibrada se muestra falso, decimos que "el modelo no se sostiene", en un momento veremos con más detalle el significado de esta expresión.

Hasta este punto

1. Si se sabe que los diferentes eventos tienen la misma probabilidad de ocurrir, entonces se puede asignar la probabilidad *a priori* que significa "antes de hacer ningún experimento".
2. Cuando no es posible conocer de antemano la probabilidad, entonces debe hacerse el experimento, que consiste en **observar** y contar las veces que sucede el evento que nos interesa. Con esta información podemos calcular **de manera aproximada** las probabilidades de cada evento, a través de las frecuencias relativas, a las que llamamos probabilidades *a posteriori*. A medida que sean más los casos observados, tanto más cerca estarán las frecuencias relativas de las probabilidades.
3. Si el supuesto inicial (que en este ejemplo es de iguales probabilidades) es verdadero, entonces las frecuencias relativas se acercarán a las probabilidades a priori.

Concepto de modelización

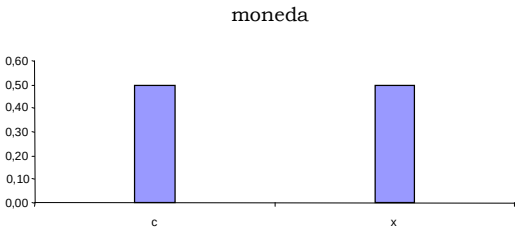
Lo que hemos hecho hasta aquí con el ejemplo elemental de la moneda es una forma muy general de tratar con los fenómenos que dependen del azar. El supuesto inicial de probabilidad ½ a cada lado, constituye un *modelo de probabilidad*, es una anticipación acerca de lo que se espera que suceda. No es arbitrario, ya que deben tenerse razones para suponer que el modelo se sostiene, pero en todos los casos es una aproximación a lo que sucede en la realidad. Tratamos de modelar (o modelizar) lo que observamos a fin de simplificarlo, pero un modelo puede ser más o menos adecuado a la realidad, por eso dijimos que cabe la posibilidad que "el modelo no se sostenga". La idea de simplificar está aquí utilizada en el sentido de elegir algunos aspectos de la realidad para construir un modelo explicativo. La riqueza y complejidad de los fenómenos sociales no se menoscaba porque se usen modelos, pero sí cuando se confunde el modelo con la realidad.

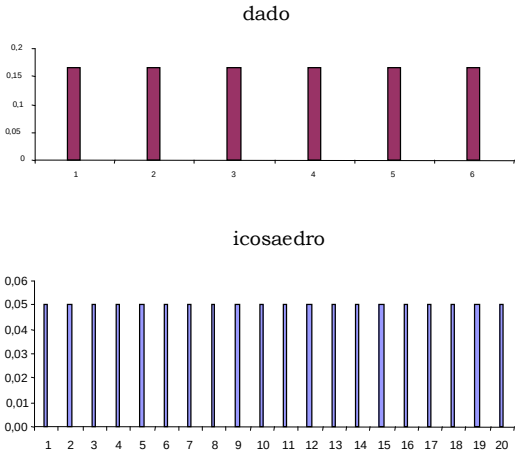
Para terminar con nuestra moneda, digamos que el modelo que resume el supuesto de iguales chances para todos los resultados, se llama "distribución uniforme", es válido también para un dado, si está equilibrado, o para cualquier fenómeno aleatorio en el que se pueda suponer que los resultados son igualmente probables. La expresión formal de ese modelo es la siguiente:

"Si un experimento aleatorio tiene distribución uniforme y k resultados posibles, entonces $P(A_i) = \frac{1}{k}$, donde A_i es uno cualesquiera de los k resultados"

Por lo tanto si se trata de una moneda (2 caras) la probabilidad de cualesquiera de sus caras será ½, si es un dado, $1/6$ y si es un icosaedro regular (veinte caras iguales, parecido a un globo de espejos) cada cara tendrá probabilidad $1/20$ de salir.

La representación gráfica de este modelo, es para cada uno de los ejemplos, la siguiente:





En estos gráficos, la idea de uniformidad (para todos los eventos la misma probabilidad) se transmite en la igual altura de todas las columnas.

Además de este modelo uniforme, en este capítulo presentaremos cuatro modelos de probabilidad que serán necesarios para lo que sigue de nuestros contenidos. Hay una gran cantidad de modelos que permiten asignar probabilidades a priori a diferentes fenómenos observables. El tratamiento que haremos a continuación será de carácter utilitario, es decir estará centrado en el uso que podemos hacer de cada distribución teórica. Haremos las referencias matemáticas mínimas que sean necesarias para comprender las propiedades y condiciones de aplicación de los modelos.

Para lo que sigue será necesario hacer más explícita una parte del vocabulario que hemos usado. Al tratar con eventos aleatorios (resultados de experimentos aleatorios), en algunos hemos usado una notación más textual, como “que la moneda salga cara” y en otros una más numérica, por ejemplo “S=4”. De aquí en adelante trabajaremos con la segunda notación que tiene muchas ventajas. La razón es que pasamos a tratar a los experimentos como observación de variables, cuyas categorías (o valores) son los resultados del experimento. En la tirada de la moneda, no es difícil cuantificar los eventos, podemos elegir cara como lado preferido y definir la variable X = “número de caras que se obtienen al lanzar una moneda equilibrada”, los valores posibles son 0 (si sale cruz) y 1 (si sale cara). En la suma de los puntos de los dos dados, S asume valores discretos del 2 al 12. La

descripción como distribución de frecuencias que hacíamos antes, se transforma ahora en distribución de probabilidades y, para los dos ejemplos citados, toma la forma:

X número de caras que se obtienen al lanzar una moneda equilibrada	P(X)
C	1/2
X	1/2
Total	1

S suma de puntos obtenidos al lanzar dos dados equilibrados	P(S)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36
Total	1

Estas variables se denominan **variables aleatorias**, porque sus valores dependen del azar y serán muy necesarias en todo lo que sigue de la materia. El aspecto de estas tablas de distribución de probabilidades es en un todo equivalente al de las de distribución de frecuencias: el total igual a uno indica la probabilidad del espacio muestral ($P(\Omega) = 1$). Sobre ellas pueden hacerse operaciones similares a las descripciones del capítulo 3.
Cuando la variable aleatoria puede asumir valores continuos, ya no es posible asignar probabilidades individualmente y debe hacerse a través de intervalos. Es el mismo problema que hallamos en el capítulo 2 y lo resolvemos del mismo modo. Se usan probabilidades acumuladas, ya sea entre valores de la variable o desde ciertos valores hacia los mayores o desde ciertos valores hacia los menores.

También es posible, sobre distribuciones de probabilidad, calcular algunas medidas descriptivas, veamos en qué se transforma la media. El procedimiento para calcularla que hemos visto consiste en multiplicar cada valor de la variable por su frecuencia y dividir por el total de casos; para las variables aleatorias eso se hace de una sola vez, multiplicando cada valor por la probabilidad y se obtiene:

$$E(x) = \sum_{i=1}^{n} x_i * P(x_i)$$

Que se llama **esperanza matemática**, la indicaremos E(x). Aunque esta expresión proviene del cálculo de la media[43], tiene una diferencia conceptual importante: la media es una medida descriptiva, por lo que describe hechos que han sido observados, puede decirse que hace referencia a algo que ya sucedió. Por el contrario la esperanza se refiere a eventos que pueden llegar a suceder, por eso tiene ese nombre, es la expectativa que tenemos sobre el devenir futuro de un experimento. Uno de los usos más difundidos de este concepto es la esperanza de vida, que mide el número promedio de años que viviría un conjunto de personas si a lo largo de toda su vida se mantuvieran las mismas tasas de mortalidad que en la actualidad (las tasas de mortalidad se usan para estimar las probabilidades de morir a cada edad). Obsérvese que es un "promedio a futuro", una descripción sobre algo que aun no sucedió, por eso se llama esperanza.

De manera análoga puede calcularse la varianza de una variable aleatoria, definida como:

$$V(x) = \sum_{i=1}^{n} (x_i - E(x))^2 * P(x_i)$$

Al referirse a una variable aleatoria, la interpretación de la varianza difiere levemente de aquella que se usó como medida descriptiva de la dispersión. Aquí indica el modo en que se alcanzará el valor de la esperanza. Una varianza grande indica un proceso aleatorio muy variable, con valores dispares que convergen lentamente al valor de la esperanza. Por el contrario, un valor pequeño de la varianza indica un proceso más estable en el que los valores difieren poco de un paso al siguiente y rápidamente se aproximan a la esperanza.

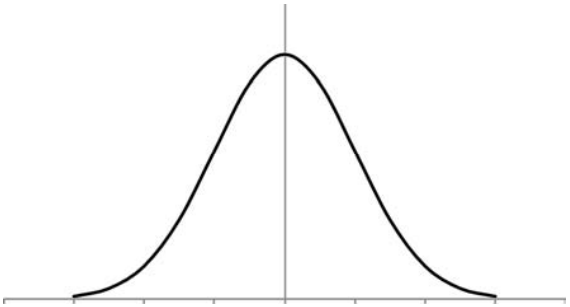
[43] Puede deducirse esto a partir del cálculo de la media, si se reemplazan las frecuencias relativas por las probabilidades de cada valor de la variable, tenemos:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * f_i}{n} = \sum_{i=1}^{n} x_i * \frac{f_i}{n} = \sum_{i=1}^{n} x_i * f_i' = \sum_{i=1}^{n} x_i * P(x_i) = E(x)$$

Modelos especiales de probabilidad

Distribución normal

La mayoría de los fenómenos naturales, sociales, psicológicos no tienen distribución uniforme, es decir, no es igualmente probable que resulte cualesquiera de los resultados. Por ejemplo, para una determinada población, el peso de los niños al nacer tiene un valor promedio y cuando un niño nace se espera que tenga un peso cercano a ese valor medio. No es igualmente probable que un niño nazca con 3500 gramos que con 5800. Son menos frecuentes los niños que nacen con pesos muy por encima o muy por debajo del promedio. De modo similar sucede con medidas psicológicas como el Cociente intelectual (IQ); se hallan con mayor frecuencia valores cercanos al promedio y la probabilidad de encontrar personas muy por encima o muy por debajo de ese promedio es menor. Para este tipo de fenómeno, hay un modelo que suele ajustar bien las probabilidades, se llama distribución normal[44] y su representación gráfica, una curva unimodal, simétrica, de forma acampanada es llamada "campana de Gauss"[45] en referencia a Johann Carl Friedrich Gauss[46].



[44] El nombre "normal" de este modelo no hace referencia al sentido coloquial del término, como juico acerca de la salud o el comportamiento humano, sino a un modelo de distribución en el que es más frecuente hallar valores cercanos al promedio y resultan igualmente infrecuentes valores extremos mayores o menores.

[45] Este modelo tiene una expresión matemática más compleja que la primera que vimos, solo a título ilustrativo la mencionamos:

$$P(z < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$$

jamás la usaremos para calcular probabilidades.

[46] 1777–1855, matemático, astrónomo y físico alemán. Aunque el primero en expresar la función y calcular áreas fue Abraham De Moivre matemático francés que trabajó en Inglaterra junto a Newton y Halley (Tankard, 1984).

Un primer elemento a tener en cuenta es que, a diferencia de los gráficos anteriores, ahora se trata de una curva con trazado continuo. Esto se debe a que esta distribución es adecuada para modelar variables continuas. Recordemos en el capítulo 2, que cuando construimos las distribuciones de frecuencia, resultaba imposible enumerar todas las categorías de una variable de este tipo, ya que son infinitas. Dijimos en ese momento que no es posible indicar la frecuencia de un valor único de una variable continua. Lo mismo vale ahora para las probabilidades: no calculamos probabilidades para valores simples de variables continuas, sí calculamos probabilidades acumuladas y, como sucedía con las frecuencias acumuladas, éstas están representadas gráficamente en el área bajo la curva.

El cálculo de las probabilidades bajo el modelo normal —o, lo que es lo mismo, de las áreas bajo la curva—, es muy complejo, por lo que se encuentra tabulado, precalculado para algunos valores de la variable. ¿De qué variable?, hemos mencionado el peso al nacer, el IQ, ¿con qué método calcularemos probabilidades para fenómenos tan disímiles? Antes que se difundiera el uso de la computadora personal, se usaban tablas, ahora usamos cualquier hoja de cálculo, donde se indica la probabilidad acumulada para diferentes valores de una variable abstracta, sin unidades, adaptable a una diversidad de fenómenos. Se trata de la variable z, que tratamos en el capítulo 3, que mide el número de desviaciones estándar —contadas desde la media—, a que se encuentra un caso individual. Una de las aplicaciones prácticas de esta variable es que permite comparar variables que miden cualidades muy diferentes. Sabemos ya que la cantidad de desviaciones estándar (z) es una medida de lo cerca o lejos que un caso se encuentra del promedio. Si la variable en estudio es adecuadamente modelada con la distribución normal, entonces podremos conocer la probabilidad de hallar casos, por ejemplo a más de dos desviaciones estándar de la media y eso tendrá una inmediata traducción a valores de la variable.

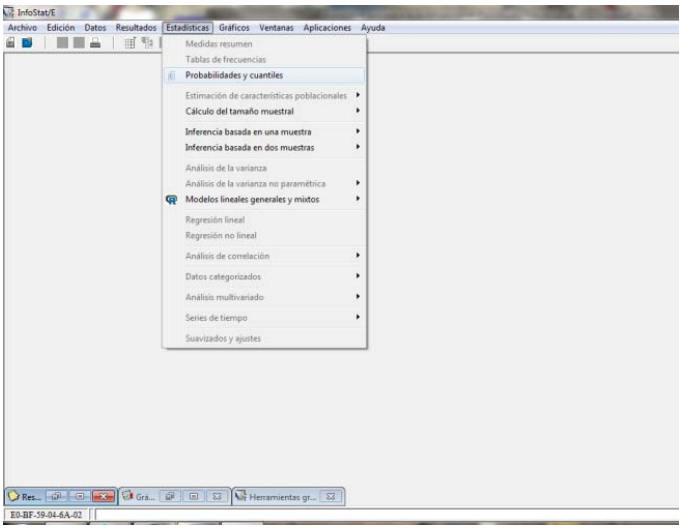
Recordemos que la variable z está definida, para un valor particular de x como:

$$z = \frac{x - \bar{x}}{s}$$

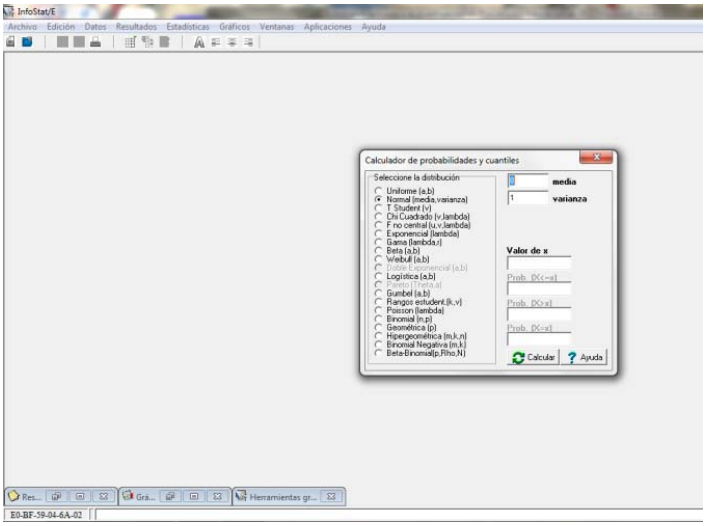
Y que tiene media igual a cero y desviación estándar igual a uno. Si x tiene una distribución normal con una media $\bar{x}$ y una desviación estándar s, entonces z tiene distribución que se llama **normal estándar**.

Dijimos que cualquier hoja de cálculo (como OpenOffice.org Calc o Excel) calcula las probabilidades (o áreas) bajo la curva normal. Nos resulta más cómodo utilizar InfoStat®, que tiene estas operaciones

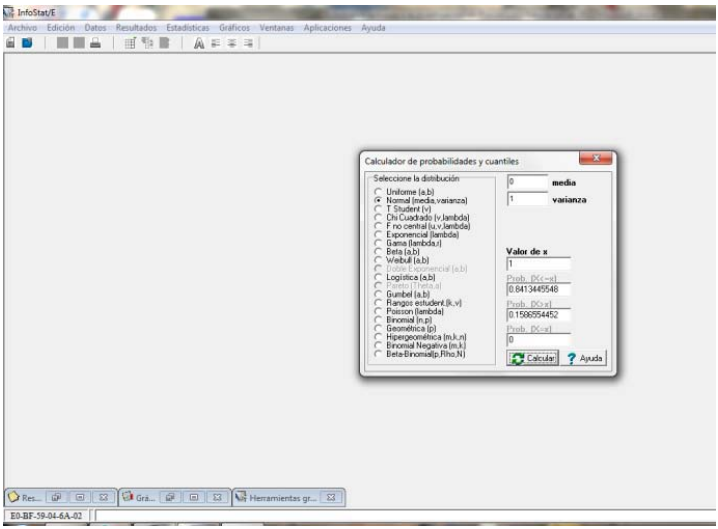
incorporadas. Sin necesidad de tener una base de datos abierta, en "Estadísticas" elegimos "probabilidades y cuantiles":



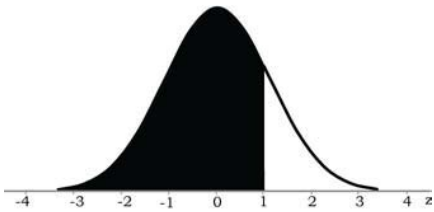
La ventana siguiente, nos pregunta por la distribución de probabilidad que usaremos, elegimos Normal (media, varianza) y por defecto, en el lugar de la media aparece un 0 y en el de la varianza un 1. Eso significa que, por defecto, trabajará con una normal estándar.



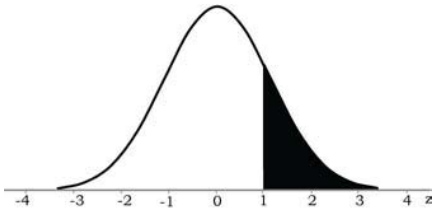
El "valor de x" es el puntaje *z* a partir del cual calcularemos la probabilidad. Vamos a introducir el valor 1 y solicitamos "calcular".



Obtenemos tres resultados expresados así (redondeados):
Prob (X<=x)=0,8413
Prob (X>x)=0,1587
Prob (X=x)=0
El primero es la probabilidad de hallar valores de *z* iguales o menores que 1. Eso significa, gráficamente, el área bajo la curva normal, desde 1 hacia la izquierda, hasta menos infinito (-∞). La representación gráfica de esta área es la siguiente:



El segundo valor es el complemento del primero (lo que le falta para llegar a 1, el área completa) y representa ahora la probabilidad de hallar valores por encima del valor que especificamos, es decir entre 1 e ∞. Gráficamente es:



Esta probabilidad podría haberse obtenido restando de 1, la probabilidad anterior, ya que: P(z<x)= 1-P(z>x)

El tercer valor dice que la probabilidad que tiene la variable de valer exactamente uno, es cero. Esto es así porque se trata de una variable continua, para la cual, como ya mencionamos, no hay probabilidades exactas. El equivalente gráfico de esta situación es que no hay área si se trata de un único valor de la variable. Al solicitar el resultado a InfoStat® para una variable continua, en ese recuadro siempre se obtiene un cero.

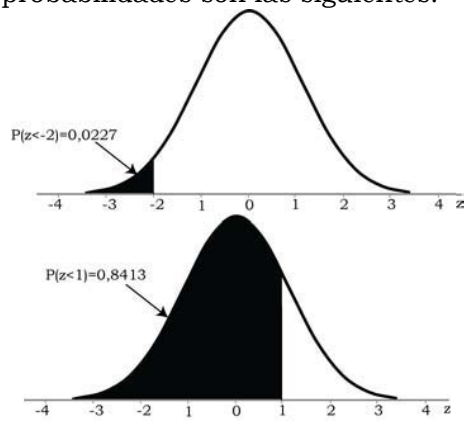
Por ahora usaremos el primero de los resultados, el que provee la probabilidad acumulada por debajo del valor que elegimos. Si solicitamos las probabilidades acumuladas por debajo de diferentes valores de *z*, empezando con los negativos, pasando por el cero, obtenemos lo siguiente:

z	*P(<z)*	Ubicación
-3,0	0,00135	
-2,5	0,00621	
-2,0	0,02275	
-1,5	0,06681	
-1,0	0,15866	
-0,5	0,30854	

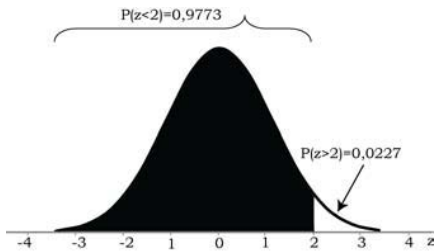
z	*P(<z)*	Ubicación
0,0	0,50000	
0,5	0,69146	
1,0	0,84134	
1,5	0,93319	
2,0	0,97725	
2,5	0,99379	
3,0	0,99865	

Las probabilidades acumuladas (es decir, las áreas a la izquierda) van creciendo desde casi cero en el valor más pequeño que pusimos (z = -3) y llegan hasta casi uno en el máximo valor (z = 3). En el modelo matemático, z tiene como campo de variación todos los valores, es decir, desde menos infinito hasta más infinito ($-\infty<z<\infty$), pero como vemos, en la realidad, los valores -4 y 4 son muy extremos, en el sentido que las probabilidades acumuladas son **casi** cero y **casi** uno respectivamente. En el gráfico, la curva se confunde con el eje para

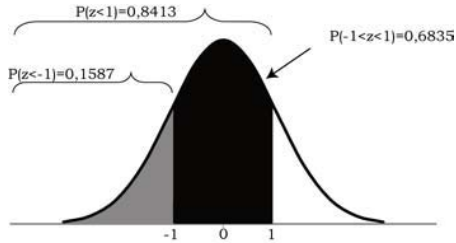
valores cercanos a 4. La notación para estos resultados es la siguiente (usando hasta cuatro decimales que es lo más frecuente), por ejemplo: P(z<-2) = 0,0227 ó también P(z<1) = 0,8413. Las representaciones gráficas de estas probabilidades son las siguientes:



A continuación vemos la representación grafica de la relación entre las probabilidades "por encima" y "por debajo" de *z=2*.



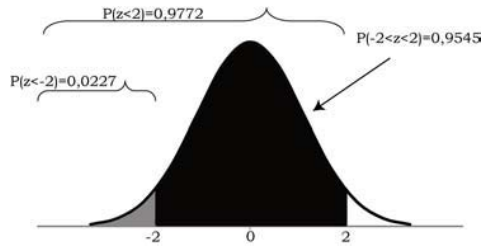
También se pueden identificar probabilidades **entre** valores de z, por ejemplo cuál es el área entre -1 y 1 (es decir cuál es la probabilidad de encontrar a z entre esos valores). Esto se escribe así: P(-1<z<1). Para calcularla solo contamos con la información sobre la probabilidad acumulada: el área por debajo de -1 vale 0,1587 y el área por debajo de 1 es 0,8413, si restamos esas dos áreas tendremos lo que queda entremedio: 0,8413 – 0,1578 = 0,6835, gráficamente:



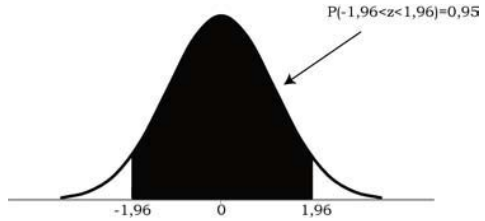
El área comprendida entre -2 y 2 se calcula del mismo modo, ya que

$$P(-2 < z < 2) = P(z < 2) - P(z < -2) = 0{,}9772 - 0{,}0227 = 0{,}9545$$

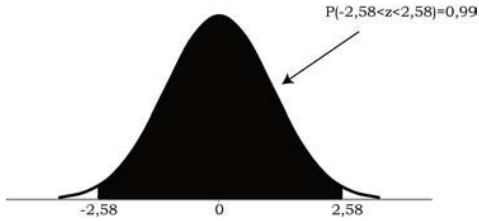
Verifique estos valores usando el procedimiento indicado con InfoStat®. La representación gráfica de esta área es:



Hay dos valores de z que conviene recordar, ya que delimitan áreas centrales que son de uso muy frecuente. El primero es el par de valores que dejan un 95% de los casos entre ellos. Según el cálculo del párrafo anterior, deben ser cercanos a -2 y 2 (±2), ya que entre ellos se ubica el 95,45% del área total. En efecto, los valores exactos son ±1,9599, que alcanza con tomarlo como ±1,96, puede verificarse en InfoStat®. Gráficamente, tenemos entonces:



El otro par de valores de interés es el que delimita un área central del 99%, es el valor ±2,5758, al que podemos recordar como ±2,58. El gráfico de esta relación entre valor z y área es:



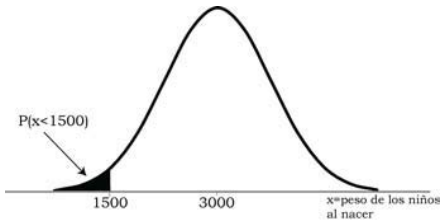
La distribución normal es muy valiosa, por muchas razones, de las cuales nos interesan dos.
La primera es que es adecuada para modelar muchos fenómenos observables, en especial variables biológicas como el *peso* de los niños al nacer, la *talla* para diferentes edades. Hay variables psicológicas como el *coeficiente intelectual*, para las que se elaboran instrumentos de medición cuyos puntajes se distribuyen de manera normal. Otras variables, como el *rendimiento en la escuela*, a menudo, pueden aproximarse de manera adecuada con un modelo normal.
La segunda razón es que cuando se extraen muestras aleatorias de una población, algunas medidas que se calculan sobre esas muestras tienen distribución normal y eso será lo que nos permita realizar nuestras primeras estimaciones de valores poblacionales a partir de muestras.

Hasta este punto trabajamos con la distribución normal de una variable abstracta, sin unidades, a la que llamamos *z*, a cuyos valores pudimos asignar probabilidades. Es en términos de z que está expresada la función que describe la curva acampanada. Pero cuando vamos a trabajar con una variable real, como las que mencionamos, por ejemplo, el peso de los niños cuando nacen o el CI, es necesario disponer de una variable concreta, a la que llamamos x. ¿Qué relación tiene esa variable, cuya distribución podría modelarse bien con una curva normal, con la *z*, que nos permite hallar las probabilidades bajo el modelo normal? Es decir, ¿qué relación tiene x con z? Ya la conocemos, porque z es el desvío estándar que definimos en el capítulo 3, es:

$$z = \frac{x - \bar{x}}{s}$$

Que cuenta la cantidad de desviaciones estándar (s) a que se encuentra un valor (x) de la media ($\bar{x}$). Para encontrar probabilidades correspondientes a valores reales de una variable y no solo los z abstractos, debemos transformar esos valores en puntajes *z*, con la expresión de arriba.

Veamos esto en un ejemplo: a partir de las historias clínicas de varios años de un hospital materno infantil, conocemos que el peso medio en el momento de nacer (de niños varones, a término, con madre entre 20 y 29 años) son 3300 gramos y que la desviación estándar es de 800 gramos. Además sabemos que la distribución normal es adecuada para describir la variable peso al nacer. Entonces podemos calcular la probabilidad que un niño al nacer tenga un peso por debajo de los 1500 gramos. El problema es averiguar el área sombreada en el siguiente gráfico:



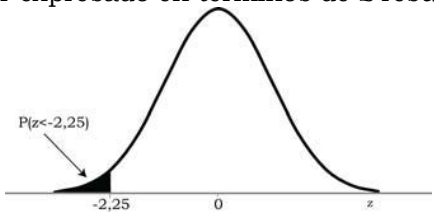
Para calcular el área marcada —que corresponde a la probabilidad solicitada— vamos a transformar el valor de x (1500) a puntaje z, haciendo:

$$z = \frac{x - \bar{x}}{s} = \frac{1500 - 3300}{800} = -2{,}25$$

Ahora es equivalente solicitar la probabilidad de x menor que 1500 que la de z menor que -2,25, en símbolos:

$$P(x < 1500) = P(z < -2{,}25)$$

y el gráfico anterior expresado en términos de z resulta:



A esta probabilidad sabemos cómo calcularla y obtenemos 0,0122.
Ésta es la probabilidad de hallar niños que al nacer tengan pesos inferiores a los 1500 gramos.

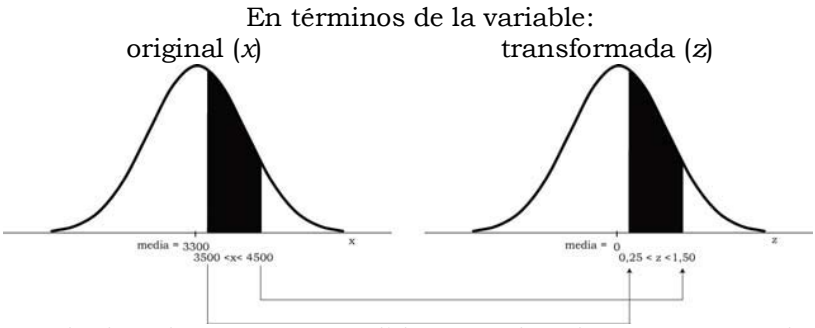
Del mismo modo podemos operar si nos interesa conocer la probabilidad de hallar niños que nazcan con peso comprendido entre 3500 y 4500 gramos. Transformando cada uno de los valores se obtiene:

$$x = 3500 \rightarrow z = \frac{3500 - 3300}{800} = 0{,}25$$
$$x = 4500 \rightarrow z = \frac{4500 - 3300}{800} = 1{,}50$$

Entonces, que x (el peso de los niños al nacer) esté entre 3500 y 4500 gramos, equivale a que z (cantidad de desviaciones estándar) se encuentre entre 0,25 y 1,50. Esto se expresa simbólicamente así:

$$P(3500 < x < 4500) = P(0{,}25 < z < 1{,}50)$$

Y se representa gráficamente así:



Para calcular el área comprendida entre los dos puntos, es decir la probabilidad de hallar a x entre 3500 y 4500, restamos, a la probabilidad acumulada hasta *z=1,50*, la acumulada hasta *z=0,25*:

$$P(3500 < x < 4500) = P(0{,}25 < z < 1{,}50) = P(z < 1{,}50) - P(z < 0{,}25)$$
$$= 0{,}9332 - 0{,}5987 = 0{,}3345$$

Leemos el resultado diciendo que la probabilidad de hallar a un niño que haya nacido con un peso comprendido entre los 3500 y los 4500 gramos es de 0,3345.

Toda variable que puede modelarse con una distribución normal puede transformarse a puntaje z y así hallar la probabilidad asociada a diferentes intervalos.

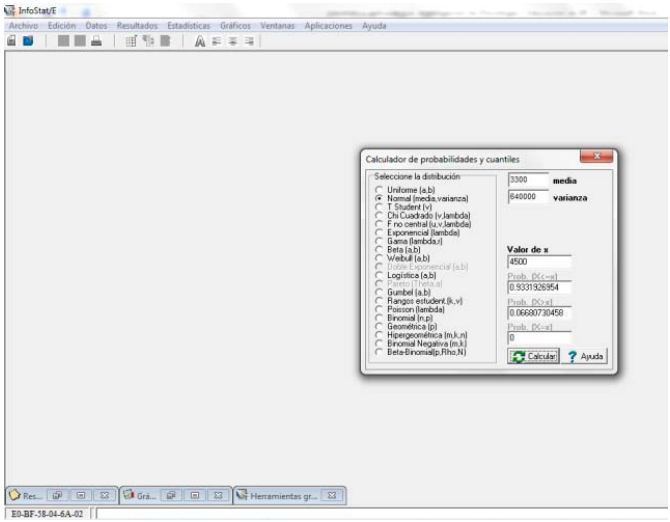
Las hojas de cálculo y, por cierto, los programas de análisis de datos incorporan esta operación, por lo que es posible ingresar directamente con los valores de la variable e informar la media y la desviación estándar, para obtener las probabilidades correspondientes. Veamos el modo de obtener el resultado de manera directa, sin pasar por z. La probabilidad que nos interesa es, como antes:

$$P(3500 < x < 4500)$$

Que se calcula como la diferencia entre la probabilidad acumulada hasta 4500 menos la acumulada hasta 3500.

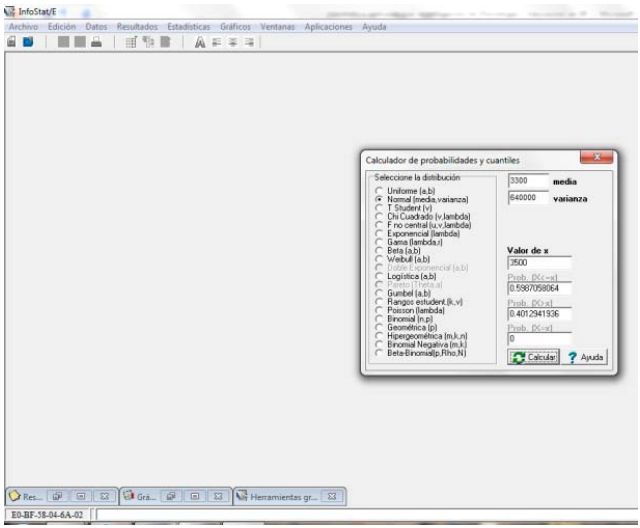
$$P(3500 < x < 4500) = P(x < 4500) - P(x < 3500)$$

Usando InfoStat®, solicitamos la distribución normal, pero ahora cambiamos la media y la varianza que aparece por defecto como cero y uno (para la variable *z*), por los valores de la media y la varianza que conocemos[47] e indicamos el primer valor de x:



La probabilidad acumulada hasta 4500 es 0,9332. Del mismo modo obtenemos la otra:

[47] Contábamos con el dato de desviación estándar igual a 800, por lo que la varianza es de 640000



Que es 0,5987, entonces

$$P(3500 < x < 4500) = P(x < 4500) - P(x < 3500) = 0{,}9332 - 0{,}5987 = 0{,}3345$$

Es el mismo resultado al que habíamos llegado a través de la transformación a puntaje *z*, pero ahora pidiéndola directamente desde *x*. Aunque éste sea el procedimiento más cómodo para calcular probabilidades, es necesario estar familiarizado con la distribución normal estándar (en *z*), por los usos que haremos en los próximos capítulos.

La idea de grados de libertad

Antes de avanzar hacia los otros modelos de probabilidad que presentaremos en este capítulo, es necesario tener una idea aproximada de un concepto que juega un papel de importancia en esas distribuciones, el de "grados de libertad", un número que, como veremos enseguida, determina la forma de algunas distribuciones. Se trata de un concepto matemático complejo, al que vamos a definir como el número de observaciones linealmente independientes que existen en una suma de cuadrados[48], pero del que haremos un uso exclusivamente instrumental. Su denominación es gl o bien df en inglés (degrees of freedom).

Consideremos el caso del cálculo de la varianza, cuyo numerador es: $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Es la suma de n términos, cada uno de ellos elevado al

[48] Solo nos interesará contar con una idea intuitiva y ejemplos de aplicación.

cuadrado, identificar los grados de libertad de esa expresión equivale a responder a ¿ccuántos términos de esa suma no dependen del valor que asumen los demás? Supongamos que se conoce la media muestral; esto establece un número fijo para la suma de las n observaciones: $x_1, x_2,\ldots x_n$, por lo que solo $n-1$ de ellas pueden elegirse "libremente", la última observación queda determinada.
Por ejemplo, si se trata de cuatro observaciones de las que sabemos que su media es cinco, entonces la suma de los cuatro números debe ser 20 ¿Cuáles son los valores posibles para las cuatro observaciones x_1, x_2, x_3, x_4? Hay infinitas combinaciones posibles que darían suma igual a 20, fijemos arbitrariamente algunos de los valores: $x_1=2, x_2=4, x_3=4$, estos tres números suman 11, por lo que el cuarto solo puede ser 9 (para lograr la suma de 20 y que entonces la media sea 5). Otra combinación posible de valores es $x_1=4, x_2=5, x_3=10$, con lo que $x_4=1$ necesariamente. Cualquiera sea la elección que se haga, siempre serán tres de los cuatro valores los que puedan elegirse libremente; decimos entonces que los grados de libertad son tres.
En general, para este problema, los grados de libertad se calcularán como $n-1$, donde n es el número de observaciones, diremos así que, de n observaciones, $n-1$ son independientes, si se fija el resultado de la suma.

Una situación diferente es cuando se trata con una tabla de doble entrada con frecuencias marginales fijas; allí los grados de libertad se interpretan de modo distinto. No están relacionados con el número de observaciones sino con la dimensión de la tabla. Veamos la siguiente tabla de dos por tres, con frecuencias marginales fijadas:

	A	B	C	Total
R				10
S				30
Total	5	10	25	40

En ella se relacionan dos variables cuyas categorías son A, B y C para la de las columnas y R y S para la de las filas. En la tabla, hemos fijado las frecuencias marginales. No hay dudas que hay infinitos valores posibles para las frecuencias conjuntas que podrían sumar lo que piden las marginales. Nos preguntamos ¿Cuántas de las frecuencias de las celdas (frecuencias conjuntas) pueden elegirse libremente? Si elegimos el valor 2 para la celda RA, queda determinado un 3 en la celda SA (para que cumpla con sumar 5 verticalmente). Si luego elegimos un 4 en la celda RB, la SB queda determinada a ser 6 (para sumar 10 en la segunda columna). Al

mismo tiempo, la celda RC debe ser también 4 para sumar 10 en la primera fila y también se determina la celda SC, que no puede ser sino 21, para sumar 30 en la segunda fila y 25 en la tercera columna. La tabla queda así completada, solo las celdas RA y RB fueron elegidas, las demás quedaron determinadas por la exigencia de respetar las frecuencias marginales.

	A	B	C	Total
R	2	4	4	10
S	3	6	21	30
Total	5	10	25	40

¿Cuántos serán los grados de libertad en este ejemplo?, de las seis celdas que debían completarse con frecuencias, solo 2 pudieron elegirse libremente. Éstos últimos son los grados de libertad: 2.
De manera general, para una tabla de dimensión *f por c* (donde *f* es la cantidad de filas y *c* la de columnas, como antes), los grados de libertad se calculan multiplicando *(f-1)* por *(c-1)*. En la tabla que acabamos de usar como ejemplo, la dimensión es 2 por 3, por lo que los grados de libertad son *(2-1)X(3-1)= 1X2=2*, que son las dos celdas cuyas frecuencias pudimos fijar "libremente".

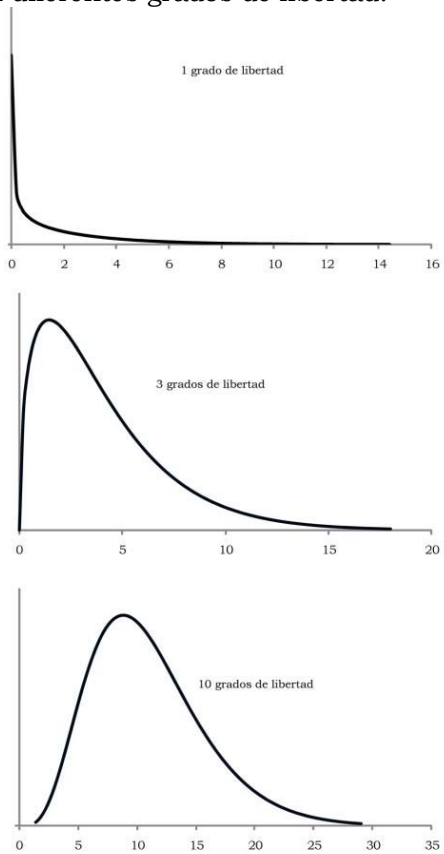
A los fines de nuestro curso, esta introducción a la idea de grados de libertad es suficiente, el concepto es amplio y su tratamiento más profundo exigiría algunos conocimientos de álgebra lineal, que no necesitamos desarrollar aquí.

La distribución ji cuadrado (χ^2)
Esta distribución es el primer modelo que veremos con forma asimétrica[49]. Una de sus aplicaciones es la estimación de la varianza, nosotros la utilizaremos más adelante, cuando tratemos sobre estadística no paramétrica, en especial para probar si una variable puede modelarse con cierta distribución (pruebas de bondad de ajuste) y también para analizar la independencia entre variables nominales.

[49] La definición de la variable χ^2 con n grados de libertad es la suma de n variables aleatorias normales estándar elevadas al cuadrado: $\chi^2 = \sum_{i=1}^{n} z_i^2$, en la que cada z es una variable distribuida normalmente con media cero y desviación estándar igual a uno.

Del mismo modo en que tratamos a la distribución normal, no haremos uso de la fórmula para calcular probabilidades, sino que las buscaremos en una tabla disponible en cualquier hoja de cálculo. Además de la forma asimétrica, hay otra diferencia con la distribución normal, la χ^2 no es una distribución única: no es suficiente especificar un valor de la variable para conocer su probabilidad acumulada sino que la forma que tenga dependerá de los recientemente mencionados "grados de libertad", los cuales en una primera aproximación tomaremos como *n-1*.
Las siguientes son las formas diferentes que tiene la distribución χ^2 para diferentes valores de los *gl*, el eje horizontal indica los valores de la variable χ^2 . Como en la normal, las probabilidades se representan como áreas bajo la curva:

Distribución χ^2 con diferentes grados de libertad:

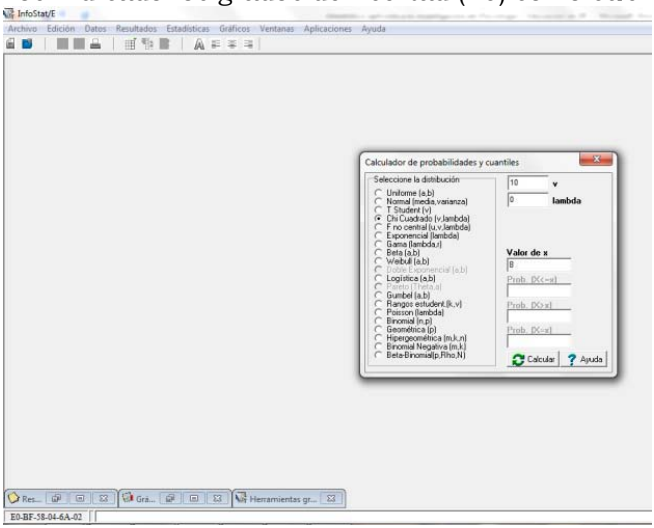


Los gráficos muestran que a medida que se incrementan los grados de libertad, la forma de la distribución gana en simetría y se asemeja a la distribución normal.

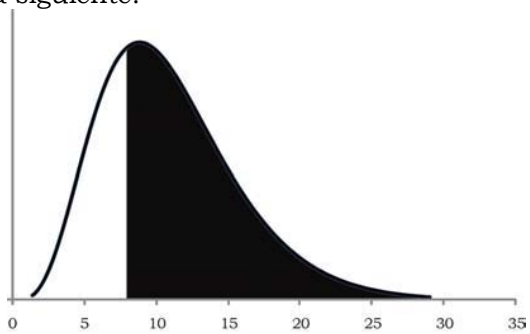
Para hallar las probabilidades correspondientes a valores de la variable, usaremos nuevamente InfoStat®. Vamos a calcular la probabilidad de hallar valores de una variable con distribución χ^2 con diez grados de libertad, superiores a 8, a lo que expresamos como:

$$P(\chi^2_{10} > 8)$$

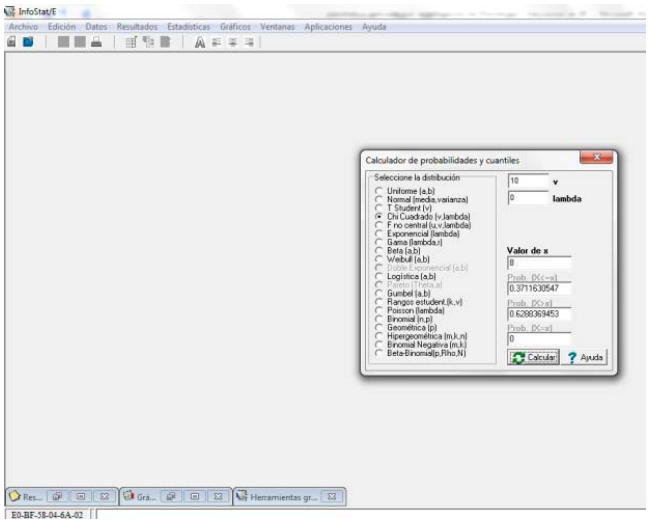
Donde hemos indicado los grados de libertad (10) como subíndice.



La letra ν (nu) indica los grados de libertad, allí colocamos el 10. Dejaremos el otro valor (lambda) en cero que es el que está elegido por defecto. El espacio señalado como "valor de x", corresponde al valor de nuestra variable, que en este caso es 8. Gráficamente el área que buscamos es la siguiente:



Y la salida InfoStat® es:



Nuevamente obtenemos dos resultados: el área por debajo (P(X<=x)) y el área por encima (P(X>x)) del valor que especificamos, las cuales son complementarias, es decir que su suma es igual a uno, que es el área completa bajo la curva. Nos interesa la segunda, por lo que:

$$P(\chi^2_{10} > 8) = 0{,}6288$$

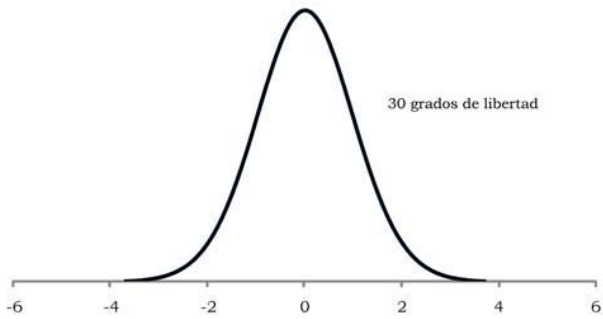
que es la superficie que está sombreada en el gráfico anterior.
La distribución χ^2 tiene muchas aplicaciones, una de las más frecuentes es para analizar la existencia de una relación entre dos variables nominales.

La distribución t de Student
El tercer modelo especial de probabilidades que nos interesa describir es la distribución t, conocida como "de Student", por el seudónimo que utilizaba quien la aplicó por primera vez, William Gosset[50].

Se trata de una distribución que, como la normal es simétrica y como la χ^2 depende de los grados de libertad[51]. Su forma es la siguiente, dependiendo de los grados de libertad:

[50] Químico y Estadístico inglés (1876-1937), era empleado de la cervecería Guiness y allí desarrolló procedimientos adecuados para trabajar con pequeñas muestras. Debió publicar con el seudónimo Student por razones de confidencialidad comercial de la empresa.

Distribución t de Student con diferentes grados de libertad:

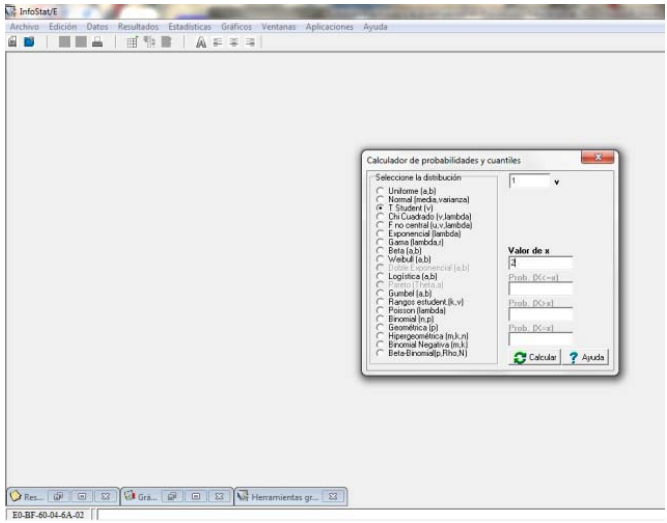


Con 30 grados de libertad, la distribución toma una forma que va haciéndose más similar a la normal. Las probabilidades acumuladas se buscan con InfoStat, donde debe informarse los grados de libertad y el valor de la variable. La probabilidad que una variable con distribución t con un grado de libertad, supere a 2 se escribe:

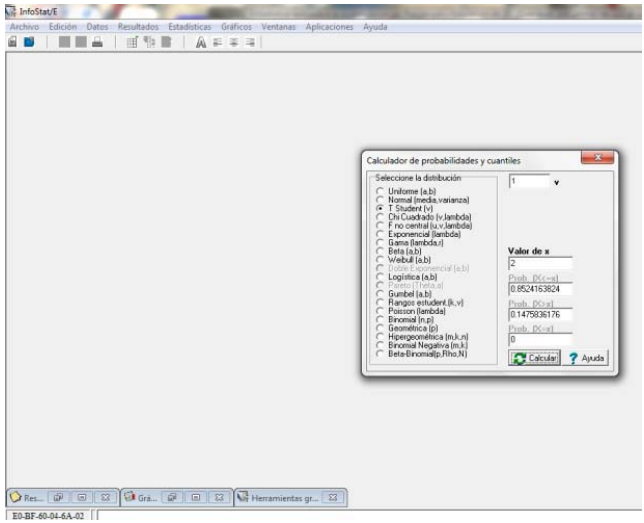
$$P(t_1 > 2)$$

[51] La definición de la variable con distribución t con n grados de libertad es el cociente entre una variable con distribución normal estándar y la raíz cuadrada de una χ^2 dividida por sus grados de libertad: $t = \frac{z}{\sqrt{\frac{\chi^2_n}{n}}}$

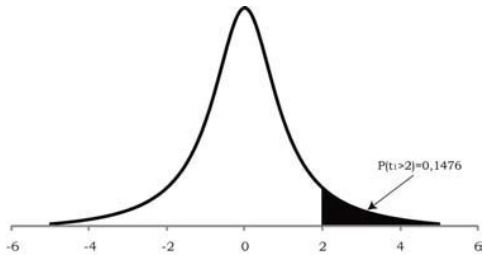
Y se solicita:



Luego de pedir “calcular” se obtiene:



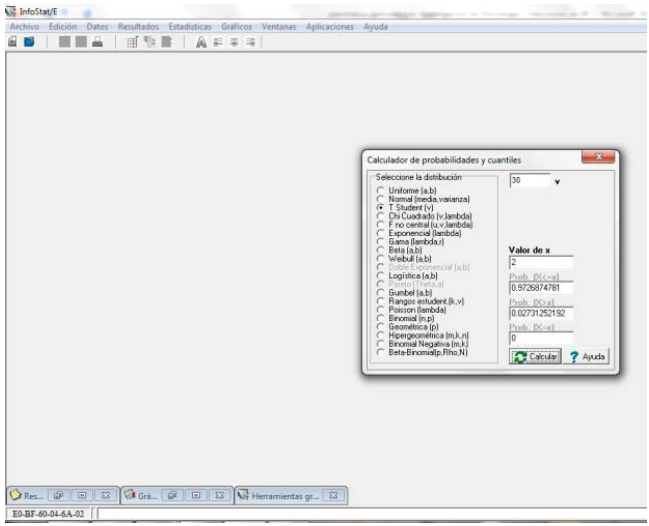
Como en las otras distribuciones, tenemos la probabilidad acumulada por debajo (Prob(X<=x)) y la probabilidad por encima (Prob(X>x)) que, cuando se suman, dan uno. El resultado que nos interesa para $P(t_1 > 2)$ es 0,1476 y se representa gráficamente así:



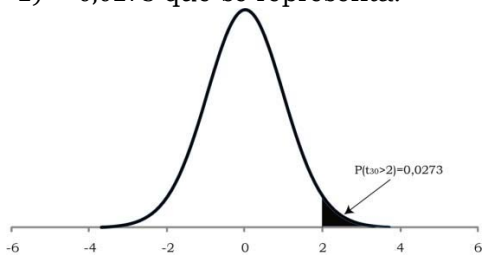
Para ver los efectos del cambio en los grados de libertad, vamos a comparar la probabilidad que una variable t con 30 grados de libertad supere a 2, es decir:

$$P(t_{30} > 2)$$

InfoStat® ofrece:



Por lo que $P(t_{30} > 2) = 0{,}0273$ que se representa:

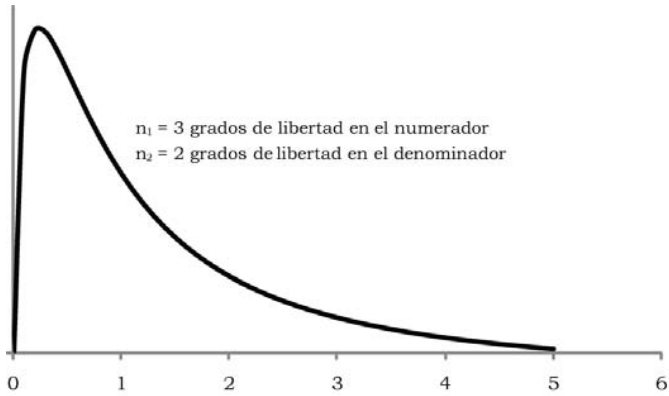


En los gráficos y en los valores de las probabilidades se ve que el aumento de los grados de libertad en la distribución t tiene el efecto de reducir la probabilidad de los valores extremos.

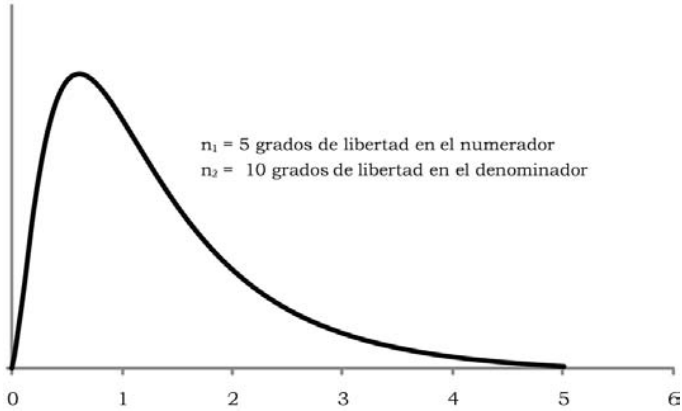
Cuando trabajemos con inferencia, veremos que la distribución t se aplica en reemplazo de la normal, cuando se trabaja con muestras pequeñas y que se va volviendo más equivalente a ella a medida que las muestra son de mayor tamaño.

La distribución F

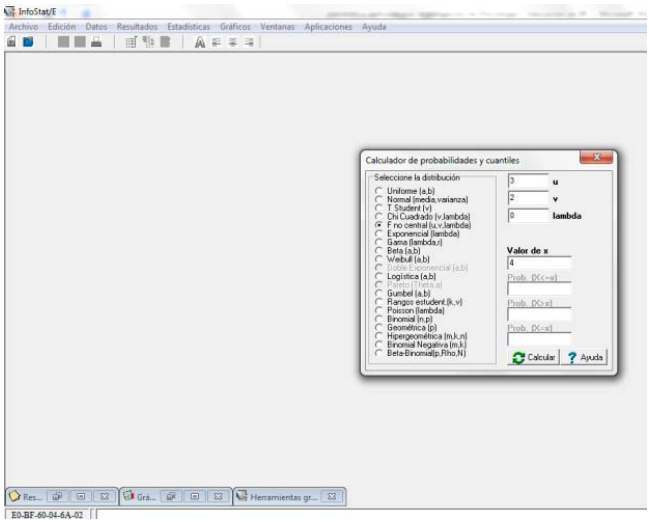
El último de los modelos de probabilidad que necesitamos para usar en los próximos capítulos en la realización de inferencias, es la distribución F de Fischer. Es una distribución asimétrica, no negativa y su forma depende de los valores de los grados de libertad del numerador y del denominador[52]. Es una curva muy asimétrica a la derecha cuando los grados de libertad son pocos y tiende a la normalidad a medida que aumentan los *gl*. Veamos dos casos de combinaciones de grados de libertad en el numerador y en el denominador:



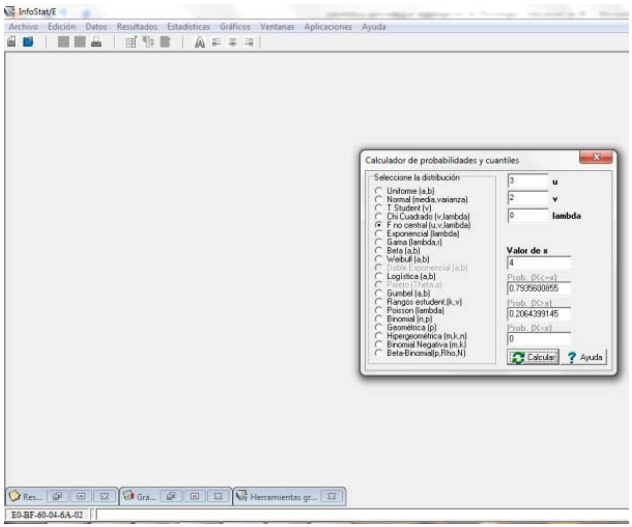
[52] Estos nombres para los grados de libertad provienen del uso que se hace de esta distribución, que es el de realizar estimaciones para cocientes de varianzas, por eso hay un numerador y un denominador. La distribución es la de una variable que es el cociente de dos distribuciones χ^2, cada una dividida por sus grados de libertad: $F_{n_1,n_2} = \frac{\chi_1^2/n_1}{\chi_2^2/n_2}$



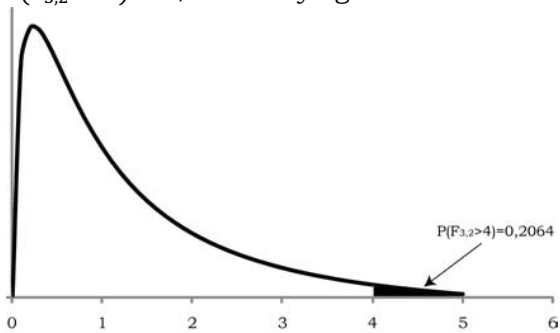
En el cálculo de las probabilidades, ahora debemos informar los grados de libertad del numerador y del denominador separadamente. Para solicitar, por ejemplo, la $P(F_{3,2} > 4)$, procedemos como antes, indicamos los *gl* en los dos espacios ofrecidos y dejamos lambda en cero:



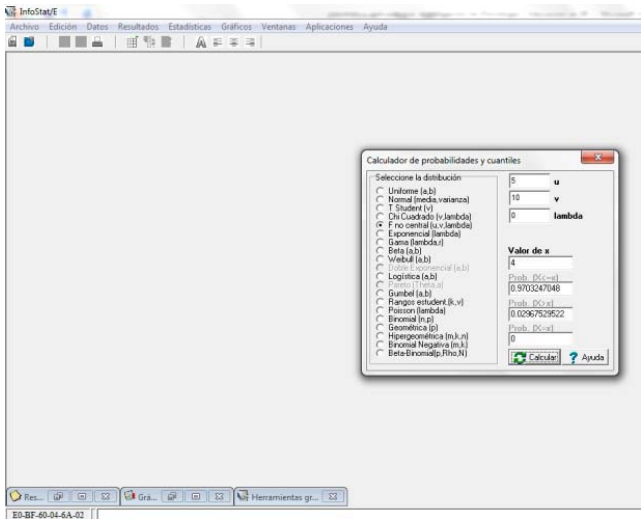
Y obtenemos:



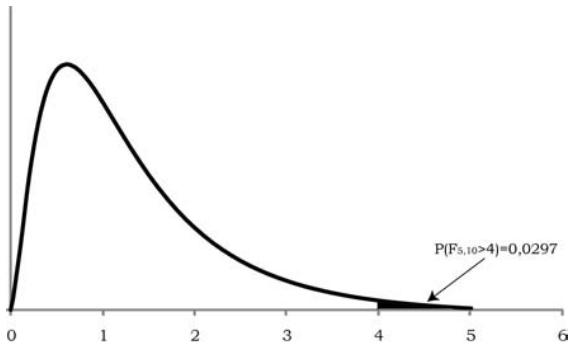
De modo que $P(F_{3,2} > 4) = 0{,}2064$. Cuyo gráfico es:



Cuando pedimos la comparación con otra variable que tenga distribución F con otros grados de libertad (5 y 10, por ejemplo), obtenemos lo siguiente:



Es decir: $P(F_{5,10} > 4) = 0{,}0297$. Esa probabilidad corresponde al área siguiente:



También en esta distribución vemos que el aumento de los *gl* tiene el efecto de reducir la probabilidad de los valores extremos. Con 3 y 2 *gl*, el valor 4 deja por encima al 20% del área total (0,2064), cuando aumentamos los *gl* a 5 y 10, ese mismo valor deja un poco menos del 3% (0,0297) por encima.

La distribución F es usada para comparar la dispersión de dos distribuciones, a través del cociente de las varianzas.

Operando con probabilidades

Con probabilidades frecuenciales

Cualquiera sea el modo a través del que se hayan asignado probabilidades a eventos, las probabilidades cumplen con ciertas propiedades generales, que trataremos a continuación y que permiten hacer operaciones con ellas.

En primer lugar, y con carácter de axiomas, las siguientes características son condición para que un número P(A) pueda ser considerado una probabilidad:

-La probabilidad es un número comprendido entre cero y uno: $0 \leq P(A) \leq 1$

-La probabilidad del conjunto completo de resultados posibles (del espacio muestral) es uno: $P(\Omega) = 1$

-La probabilidad de la unión de dos eventos que se excluyen mutuamente es la suma de las probabilidades de cada uno de ellos: $P(A \cup B) = P(A) + P(B)$

La definición frecuencial (a posteriori), así como todos los modelos de asignación de probabilidad a priori que mencionamos cumplen con estas condiciones.

A fin de ver con más claridad el uso y las aplicaciones de estas exigencias, pasemos a una distribución conjunta de dos variables con asignación de probabilidades frecuenciales, es decir empíricas.

Se trata de la relación entre la ciudad donde se vive y la intención de voto. Las categorías de la ciudad son: Córdoba, Rosario y Mendoza. Los partidos políticos son cuatro y los llamaremos Q, R, S y T. Supongamos que las siguientes son las frecuencias observadas luego de recoger los datos:

Tabla 2: Distribución del partido al que declara que va a votar y la ciudad de residencia.

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	50	650
	Rosario	100	150	60	70	380
	Mendoza	50	150	100	200	500
Total		350	600	260	320	1530

Calculemos algunas probabilidades a partir de las frecuencias relativas.

Probabilidades marginales
Cuando se consideran las categorías de una variable sin tener en cuenta a la otra, usamos las frecuencias de los márgenes de la tabla, esas son las llamadas **frecuencias marginales**. La probabilidad que una persona elegida al azar viva en Córdoba (sin importar a qué partido piense votar) es 650/1530. De manera equivalente, la probabilidad de encontrar por azar a alguien que piense votar al partido S (cualquiera sea su ciudad) es 260/1530. Las escribimos simplemente P(Córdoba) y P(S) respectivamente. En las tablas siguientes destacamos las frecuencias que participan en el cálculo de estas probabilidades.

Tabla 3: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad marginal P(Córdoba)

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	50	**650**
	Rosario	100	150	60	70	380
	Mendoza	50	150	100	200	500
Total		350	600	260	320	1530

Tabla 4: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad marginal P(S)

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	50	650
	Rosario	100	150	60	70	380
	Mendoza	50	150	100	200	500
Total		350	600	**260**	320	1530

Probabilidades conjuntas o de la intersección de eventos
Las usamos para hallar la probabilidad de ocurrencia simultánea de una categoría de cada variable. Por ejemplo ¿Cuál es la probabilidad de encontrar por azar a alguien que tenga viva en Rosario años **y** que piense votar al partido R? La cantidad de individuos que cumplen **simultáneamente** las dos condiciones es de 150, por lo que la probabilidad se calcula como 150/1530. Hemos destacado la conjunción "y", junto al "simultáneamente" porque en este caso se

piden dos condiciones juntas. Por eso estas son llamadas **probabilidades conjuntas**.
En teoría de conjuntos, corresponden a la intersección de dos conjuntos, que se indica con el signo ∩, por lo que el evento "vivir en Rosario" y al mismo tiempo "decir que se va a votar a R", se escribe "$Rosario \cap R$".
Esa intersección puede verse gráficamente en el cruce de la fila con la columna correspondiente.

Tabla 5: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad conjunta P(Rosario y R)

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	50	650
	Rosario	100	**150**	60	70	380
	Mendoza	50	150	100	200	500
Total		350	600	260	320	1530

¿Qué sucede si aplicamos esta operación a dos eventos que corresponden a dos categorías de la misma variable?, por ejemplo, ¿cuál es la probabilidad de encontrar a alguien que diga que votará a Q y a R? Esos eventos no pueden suceder juntos porque son incompatibles: solo uno de los dos puede suceder. La intersección entre ellos es imposible, por lo que la probabilidad es cero. Es el mismo caso de buscar a alguien que viva en Córdoba y también en Mendoza, es claro que no hay intersección entre estos conjuntos; dicho de otra forma, la intersección es el conjunto vacío. A estos eventos que no pueden suceder simultáneamente, los llamaremos mutuamente excluyentes. Como recordamos, esa es la condición que deben cumplir las categorías de cualquier variable.
Si dos eventos son mutuamente excluyentes entonces, su probabilidad conjunta es cero. En lenguaje de conjuntos:

$$si\ A \cap B = \emptyset\ entonces\ P(A \cap B) = 0$$

Probabilidad de la unión de eventos mutuamente excluyentes
Estas probabilidades sirven para analizar la ocurrencia de uno u otro de dos eventos, cuando éstos no pueden suceder simultáneamente. Por ejemplo: ¿qué probabilidad hay de encontrar a alguien que piense votar a Q **o** a R? Esto quiere decir "una cosa o la otra", se trata de una disyunción, es decir, la unión de los dos eventos. En el lenguaje de la

teoría de conjuntos la unión de dos conjuntos se indica con el símbolo ∪, por lo que decir "A o B" equivale a decir "A∪ B".
El total de quienes cumplen con la condición de votar a Q o a R (sin tener en cuenta la ciudad) es de 950 (350 + 600), por lo que la probabilidad es de 950/1530.

Tabla 6: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad P(Q o R) (eventos disjuntos).

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	50	650
	Rosario	100	150	60	70	380
	Mendoza	50	150	100	200	500
Total		**350**	**600**	260	320	1530

De modo equivalente, la probabilidad de seleccionar al azar a alguien que viva en Córdoba o en Rosario es de 1030/1530, donde hemos sumado las dos primeras categorías (Córdoba y Rosario).

Tabla 7: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad P(Córdoba ó Rosario) (eventos disjuntos).

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	50	**650**
	Rosario	100	150	60	70	**380**
	Mendoza	50	150	100	200	500
Total		350	600	260	320	1530

En estas probabilidades, admitimos que se cumpla cualquiera de las dos condiciones (Q **o** R en el primer caso, Córdoba **o** Rosario en el segundo). En los dos ejemplos se trata de eventos que no pueden suceder simultáneamente, por ser categorías de una de las variables, son mutuamente excluyentes, por lo que la probabilidad de su ocurrencia conjunta es cero. En estos casos, la probabilidad de la unión es simplemente la suma de las probabilidades de los dos eventos: $P(A \cup B) = P(A) + P(B)$

Aplicada a los ejemplos:
$P(Q \cup R) = P(Q) + P(R)$

$$P(Córdoba \cup Rosario) = P(Córdoba) + P(Rosario)$$

Probabilidad de la unión de eventos no mutuamente excluyentes
Vamos ahora a incluir en la operación de unión de eventos, aquellos que no se excluyan mutuamente. Para el ejemplo con el que venimos trabajando, cambiamos las condiciones de esta unión de eventos: Ahora preguntamos: ¿Cuál es la probabilidad de hallar por azar a alguien que viva en Córdoba **o** que piense votar al partido T? Otra vez es una disyunción, por lo que admitimos cualquiera de los dos eventos: que viva en Córdoba (sin importar a quién piense votar) **o** que piense votar a T (cualquiera sea su ciudad). Si intentamos el mismo procedimiento que en el caso anterior, deberíamos sumar las probabilidades: 650/1530 + 320/1530, con solo observar la Tabla 8, vemos que los 50 individuos que cumplen las dos condiciones (viven en Córdoba y votarán a T), han sido contados dos veces: en los 650 y en los 320, por lo que deben descontarse del resultado haciendo 650/1530 + 320/1530 – 50/1530.
¿Por qué sucedió esto?, porque los eventos cuya unión estamos considerando pueden ocurrir simultáneamente, tienen intersección y es esa intersección justamente la que aparece en el cálculo de las dos probabilidades que se suman. Esta última expresión tiene forma:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Y es la expresión más general para el cálculo de la probabilidad de la unión de conjuntos. Esta fórmula toma la forma simplificada que usamos antes $P(A \cup B) = P(A) + P(B)$ solo cuando A y B son disjuntos, es decir cuando se excluyen mutuamente como lo indica el tercer axioma.
Aplicada al ejemplo, el cálculo es:

$$P(Córdoba \cup T) = P(Córdoba) + P(T) - P(Córdoba \cap T)$$
$$= \frac{650}{1530} + \frac{320}{1530} - \frac{50}{1350} = \frac{920}{1530} = 0{,}60$$

Tabla 8: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad P(Córdoba o T) (eventos no disjuntos).

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	***50***	**650**
	Rosario	100	150	60	70	380
	Mendoza	50	150	100	200	500
Total		350	600	260	**320**	1530

Probabilidad condicional

Este es el caso en que necesitamos calcular una probabilidad bajo una condición, que restringe el conjunto de resultado posibles. Se aplica cuando se cuenta con información adicional antes de calcular una probabilidad, por ejemplo que se sepa que la persona seleccionada al azar vive en Córdoba. ¿Cuál es la probabilidad que piense votar al partido S? El planteo es tal que preguntamos cuál es la probabilidad de votar a S si se sabe que vive en Córdoba. Vivir en Córdoba es la condición y se escribe: S/Córdoba, y lo leemos "S, dado que vive en Córdoba".

En este caso, el dato "vive en Córdoba" es una restricción sobre el conjunto total, ya no debemos tener en cuenta a las 1530 personas del total, sino solo a los que cumplen con la condición de vivir en Córdoba. Entonces ahora, el nuevo total es de solo 650 personas, los que viven en Córdoba. De ellos, 100 piensan votar a S, por lo que la probabilidad que nos interesa es 100/650.

Entonces: $P\left({}^{S}/_{Córdoba}\right) = \frac{100}{650} = 0{,}15$.

Tabla 9: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad condicional P(*S/Córdoba*)

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	**100**	50	650
	Rosario	100	150	60	70	380
	Mendoza	50	150	100	200	500
Total		350	600	260	320	1530

Razonando del mismo modo, si se sabe que la persona elegida piensa votar a R, el total queda restringido a 600 casos (los que cumplen con esa condición). Si nos interesa la probabilidad que viva en Mendoza, bajo esa restricción resulta: 150/600.

Entonces: $P\left({}^{Mendoza}/_{R}\right) = \frac{150}{600} = 0{,}25$

Tabla 10: Distribución del partido al que declara que va a votar y la ciudad de residencia. Esquema para el cálculo de la probabilidad condicional P(*Mendoza/R*)

		Partido al que dice que votará				Total
		Q	R	S	T	
Ciudad	Córdoba	200	300	100	50	650
	Rosario	100	150	60	70	380
	Mendoza	50	**150**	100	200	500
Total		350	600	260	320	1530

En estos dos últimos ejemplos, el cambio respecto de todos los anteriores es que el denominador de las probabilidades ya no es 1530 sino un número menor, que resulta de haber impuesto previamente una condición: que viva en Córdoba en el primer caso y que haya votado a R en el segundo.

Como se ve, estas probabilidades condicionales no son conmutativas:

-Se selecciona una persona al azar entre quienes votarán a R, ¿Cuál es la probabilidad que viva en Mendoza? Se escribe P(*Mendoza/R*) y vale 150/600

-Se selecciona una persona al azar entre los que viven en Mendoza, ¿Cuál es la probabilidad que vaya a votar a R? Se escribe P(*R/Mendoza*) y vale150/500.

Relación entre probabilidades condicionales y conjuntas
Compararemos ahora la probabilidad de hallar alguien que vaya a votar a Q y que viva en Córdoba (P(*Q∩Córdoba*)) con la probabilidad que vaya a votarlo si se sabe que vive en Córdoba (P(*Q/Córdoba*)).

$$P(Q \cap Córdoba) = \frac{200}{1503}$$

$$P(Q/Córdoba) = \frac{200}{650}$$

Si dividimos entre sí estas dos expresiones obtenemos:

$$\frac{P(Q \cap Córdoba)}{P(Q/Córdoba)} = \frac{\frac{200}{1503}}{\frac{200}{650}} = \frac{650}{1503}$$

El último cociente es la probabilidad marginal correspondiente a Córdoba, por lo que:

$$\frac{P(Q \cap Córdoba)}{P(Q/Córdoba)} = P(Córdoba)$$

Esta expresión, que es general, nos ofrece una relación muy útil entre las probabilidades condicional y conjunta. Una forma más frecuente de escribir esta relación, para dos eventos cualesquiera A y B es:

$$P(A \cap B) = P(B) * P\left(A/B\right)$$

Si escribimos la intersección en orden inverso[53], tenemos:

$$P(B \cap A) = P(A) * P\left(B/A\right)$$

Como son iguales los primeros miembros de las dos expresiones anteriores, igualamos los segundos miembros, para obtener:

$$P(B) * P\left(A/B\right) = P(A) * P\left(B/A\right)$$

Esta igualdad relaciona las probabilidades condicionales en un orden o en el otro. Conociendo las probabilidades de A y B, esa igualdad nos permite pasar de P(A/B) a P(B/A), veremos más adelante que se trata de un resultado muy valioso.

Con probabilidades a priori

Veamos el uso de estas operaciones con probabilidades usando ahora un experimento en el que asignamos probabilidades a priori. Sea una caja que contiene 4 fichas rojas y 3 azules. ¿Cuál es la probabilidad de sacar una roja en la primera extracción? Como #Roja es 4 y #Ω es 7, la probabilidad vale 4/7. De mismo modo, la probabilidad de una azul es 3/7.

Hagamos ahora dos extracciones sucesivas de modo tal que no reponemos la primera ficha antes de sacar la segunda, este tipo de extracción se llama *sin reposición*. Saco la primera, veo su color y saco la segunda sin devolver la primera. ¿Cuál es la probabilidad que salga la segunda azul si la primera fue roja? En este caso, a la segunda extracción hay 3 azules sobre un total de 6 fichas (porque ya sacamos una), entonces #Azul = 3 y #Ω = 6 y la probabilidad es $P(A_2/R_1) = 3/6$ (ó 1/2).

Otro caso: ¿Cuál es la probabilidad que la segunda sea azul si la primera fue azul? Ahora quedan 2 azules, porque ya sacamos una, sobre un total de 6 fichas, por lo que la probabilidad es $P(A_2/A_1)=2/6$ (ó 1/3). Sucede entonces que la probabilidad de obtener una ficha

[53] La intersección de dos eventos es conmutativa: P(A∩B)=P(B∩A)

azul a la segunda extracción *depende* de lo que haya resultado en la primera.
Otro problema: ¿cuál es la probabilidad de sacar dos rojas en dos extracciones sin reposición? Lo escribimos: $P(R_1 y R_2)=P(R_1 \cap R_2)$
Y aplicamos la relación que encontramos al final del apartado anterior:
$P(R_1 \cap R_2)=P(R_1)*P(R_2/R_1)=(4/7)*(3/6)=2/7$

Que es la probabilidad que salga roja la primera multiplicada por la probabilidad que salga roja la segunda condicionada a que ya haya salido roja la primera (es decir, con el espacio muestral restringido) .

Del mismo modo, la probabilidad de sacar dos azules es
$P(A_1 \cap A_2)=P(A_1)*P(A_2/A_1)=(3/7)*(2/6)=1/7$

Nuevamente, es la probabilidad de azul la primera por la probabilidad de azul la segunda condicionada a que ya haya salido azul la primera.
Ahora preguntamos por la probabilidad de obtener una azul y una roja, en cualquier orden. Esto equivale a pedir azul la primera y roja la segunda o bien roja la primera y azul la segunda, por lo que:

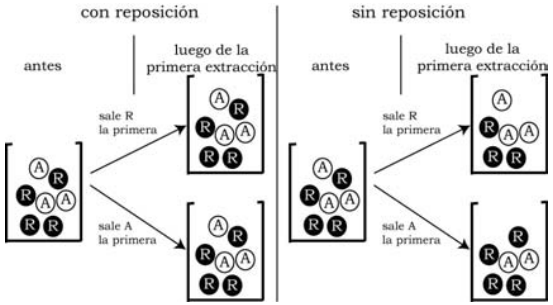
$$P((A_1 \cap R_2) U (R_1 \cap A_2))= P(A_1 \cap R_2)+P(R_1 \cap A_2)=P(A_1)P(R_2/A_1)+P(R_1)*P(A_2/R_1)$$

Si cambiamos el experimento reponiendo ahora la primera ficha antes de extraer la segunda, llamamos al experimento *extracción con reposición*. De este modo se restaura el espacio muestral al estado inicial. Por esta razón, la probabilidad de la segunda extracción será la misma que la de la primera para cualquier evento. Por ejemplo, $P(A_2/A_1)$ es $3/7$, como lo es también $P(A_2/R_1)$. Que la primera haya sido azul o roja no afecta la probabilidad de la segunda extracción, ya que se la repone: la segunda extracción no depende de la primera. En el caso de extracciones con reposición —en que la segunda extracción no se ve afectada por el resultado que haya dado la primera— decimos que los eventos son **independientes** y resulta que, para dos eventos cualesquiera A y B:

$$P(A/B) = P(A)$$

Lo cual, dicho en lenguaje cotidiano nos indica que, para la ocurrencia del evento A, no importa que haya o no sucedido el evento B. Debido a esto, en nuestro ejemplo, la probabilidad de obtener dos rojas es:
$P(R1 \cap R2)=P(R_1)*P(R_2)$

Esquema 1: Efectos de extraer una ficha sin reponerla o reponiéndola, en las condiciones que quedan para la segunda extracción



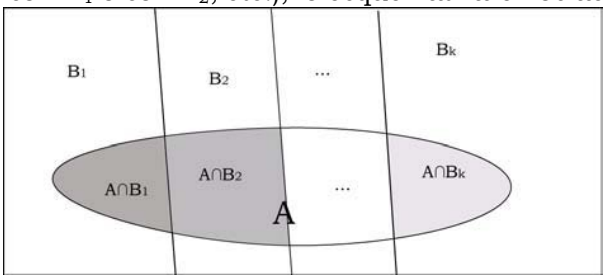
Este concepto de independencia entre eventos en muy valioso para analizar uno de nuestros más importantes problemas: las relaciones entre variables. Veamos su aplicación a la Tabla 2. Si la intención de voto fuera independiente de la ciudad donde se vive (quiere decir si votaran del mismo modo personas de las diferentes ciudades) la probabilidad de encontrar una persona que vote a R si vive en Córdoba sería simplemente la probabilidad de votar a R, es decir P(*R/Córdoba*)=P(*R*) y del mismo modo para los demás eventos. En nuestro ejemplo no se obtiene esa igualdad, ya que
P(*R/Córdoba*)=300/650=0,46, mientras que:
P(*R*)=600/1530=0,39

Por lo que estos eventos no son independientes en sentido estadístico.
En este punto conviene regresar al capítulo "Relaciones entre variables" y observar el modo en que se discutió este problema, cuando se calcularon las frecuencias que se esperarían si las variables fueran independientes. Allí encontramos que dos variables son independientes si la frecuencia relativa de cada celda resulta del producto de las frecuencias relativas marginales que le corresponden. En el lenguaje de las probabilidades, encontramos ahora el mismo resultado y lo expresamos diciendo que si los eventos A y B son independientes, entonces P(A∩B)=P(A)*P(B).

Una consecuencia importante de las probabilidades condicionales: El teorema de Bayes

La aplicación que presentamos en este último apartado usa probabilidades condicionales para deducir la probabilidad que tiene un evento observado de provenir de diferentes eventos previos. Por esta razón se denomina también "teorema de las causas"[54]. Nos interesa porque es un resultado que permite "aprender de la experiencia", lo que quiere decir que da los medios para usar la información disponible para modificar las probabilidades de determinados eventos. Tiene mucho valor en Ciencias de la Salud, porque es frecuente conocer cuál es la probabilidad a priori que un paciente tenga determinada patología (la prevalencia de la enfermedad) pero, una vez que se dispone de indicadores clínicos, esa probabilidad cambia. De manera equivalente, si se conoce cuál es la probabilidad que un alumno termine una carrera universitaria, esa probabilidad puede modificarse, una vez que se cuenta con información adicional, como el número de materias que ya ha aprobado.

De manera general, si B_1, B_2...B_k son eventos mutuamente excluyentes que completan el espacio muestral (es decir que la unión de todos esos eventos es Ω), y un evento A puede suceder en intersección con cualquiera de ellos (es decir que A puede suceder en intersección con B_1 ó con B_2, etc.), lo esquematizaremos así:



Por ejemplo, los eventos B pueden ser los diferentes tipos de escuela secundaria de la que provienen los alumnos y el evento A es "terminar la carrera". Hay quienes terminan la carrera y quienes no lo hacen, y tanto unos como otros pueden provenir de escuelas de cualesquiera de los tipos B_1, B_2, etc. La primera de las intersecciones representa a alumnos que cumplen simultáneamente A (haber terminado la carrera) y B_1 (provenir de una escuela del tipo que ese grupo define).

[54] Fue enunciado por Thomas Bayes (1702-1761), matemático inglés.

Otro ejemplo es que los eventos B sean solo dos:
B_1: tener una determinada enfermedad
B_2: no tener esa enfermedad

Si el evento A es "una prueba diagnóstica dio positiva", tenemos intersecciones que corresponden a las personas a las que la prueba dio positiva y tienen la enfermedad ($A \cap B_1$) y a aquellas a quienes la prueba dio positiva pero no la tienen ($A \cap B_2$). La segunda intersección corresponde a los casos llamados "falso positivo", enseguida volveremos sobre ellos.

Con esta notación y esa relación entre los eventos B_1, B_2…B_k y A, el teorema de Bayes se expresa de la siguiente manera:

$$P\left(B_j/A\right) = \frac{P(B_j)P\left(A/B_j\right)}{\sum_{i=1}^{k} P(B_i)P\left(A/B_i\right)}$$

El valor de este teorema es que permite pasar de la probabilidad simple de uno de los eventos B (en general, del que llamamos B_j), a la probabilidad "corregida", a partir de la información que aporta el evento A. Si se conoce inicialmente la probabilidad del evento B_j, el teorema permite calcular la probabilidad de B_j, luego de haber agregado la condición A.

Veamos un ejemplo sencillo: se dispone de dos frascos, el primero de ellos tiene 20 caramelos de menta y 10 de frutilla, el segundo contiene 20 de menta y 20 de frutilla. Se elije un frasco al azar y luego se extrae de él un caramelo, que resulta ser de menta, nos preguntamos por la probabilidad que el caramelo provenga del primer frasco. En ausencia de toda información, los dos frascos son igualmente probables, por lo que la probabilidad de cada uno es 0,50: P(F1)=0,50 y P(F2)=0,50. Si el caramelo fue extraído del primer frasco, la probabilidad de que sea de menta es: P(M/F1)=20/30=0,67, mientras que si proviene del segundo frasco es P(M/F2)=20/40=0,50. La pregunta es por P(F1/M), debemos invertir una probabilidad condicional, por lo que usaremos el teorema de Bayes:

$$P(F_1/M) = \frac{P(M/F_1) * P(F_1)}{P(M/F_1) * P(F_1) + P(M/F_2) * P(F_2)}$$

Reemplazando, tenemos:

$$P(F_1/M) = \frac{0{,}67 * 0{,}50}{0{,}67 * 0{,}50 + 0{,}50 * 0{,}50} = \frac{0{,}33}{0{,}58} = 0{,}57$$

Este resultado dice que, con el dato que el caramelo extraído es de menta, corregimos la probabilidad de provenir del primer frasco, que a priori era de 0,50; a 0,57. En este sentido la fórmula de Bayes nos permite usar la información para corregir probabilidades a priori.
En el ejemplo sobre la enfermedad y su diagnóstico, se dispone inicialmente de la probabilidad que tiene una persona cualquiera de padecer la enfermedad, esa es $P(B_1)$, luego esa probabilidad cambia cuando se agrega el dato que dice que a la persona la prueba le dio positiva.
Veamos un ejemplo que presentó Cohen (1994) en un artículo crítico hacia los procedimientos tradicionales de análisis estadístico. La aplicación ilustra el aporte del teorema a las interpretaciones de los resultados que arrojan las pruebas diagnósticas.
La prevalencia de esquizofrenia en adultos es de aproximadamente el 2%, que indica que aproximadamente 2 de cada 100 personas en la población general de adultos padece la enfermedad. Se dispone de un conjunto de pruebas diagnósticas del que se estima que tiene al menos un 95% de precisión al hacer diagnósticos positivos (sensibilidad) y aproximadamente 97% de precisión al declarar normalidad (especificidad).
Para expresar formalmente estos datos, tratamos por un lado, la situación real, la de ser esquizofrénico o no serlo. Llamamos E al evento "el paciente es esquizofrénico" y noE al evento "el paciente no es esquizofrénico". Por lo que, elegida una persona al azar, su probabilidad de ser esquizofrénico es $P(E)=0,02$, la probabilidad que no lo sea es $P(noE)=0,98$.
Por otro lado tenemos el resultado del conjunto de pruebas, que pueden dar positivas o negativas. La sensibilidad se escribe así: $P(+/E)=0,95$, que quiere decir que, aplicada a sujetos esquizofrénicos, el 95% de las veces la prueba dará un resultado positivo, que conducirá al diagnóstico correcto de la enfermedad. El complemento de esa probabilidad, 5%, es la probabilidad de dar un resultado negativo ante un caso de alguien que sí es esquizofrénico, se denomina resultado "falso negativo" y solo puede identificarse ante pruebas posteriores más sensibles o por el desarrollo de otros síntomas, que dan más elementos para realizar el diagnóstico. Escribimos entonces que $P(-/E)=0,05$.
Ante personas que no son esquizofrénicas, la prueba da, en el 97% de los casos resultado negativo (correctamente), es decir: $P(-/noE)=0,97$. Su complemento, del 3%, es la probabilidad de hallar un resultado

positivo en alguien que no es esquizofrénico[55], se denomina "falso positivo" y su probabilidad se escribe: *P(+/noE)=0,03*.

Dado un paciente cuyas pruebas dan un resultado positivo, nos preguntamos por la probabilidad que efectivamente sea esquizofrénico. Antes de conocer la respuesta al problema piénselo un momento por su cuenta y ofrezca un valor aproximado para esa probabilidad.

Este es un problema que requiere que se invierta una probabilidad condicional, ya que conocemos la probabilidad de obtener un resultado positivo si el individuo es esquizofrénico *P(+/E)*, y queremos saber la probabilidad que sea esquizofrénico dado que la prueba dio positiva, que es *P(E/+)*.
En la aplicación del teorema de Bayes, el universo está compuesto por un 98% de no esquizofrénicos y un 2% de esquizofrénicos y la prueba puede dar positiva tratándose de alguien enfermo (muy frecuentemente) o estando sano (con poca probabilidad). Reemplazamos en la expresión del teorema de Bayes y tenemos:

$$P\left(E/_{+}\right) = \frac{P\left(+/_{E}\right) * P(E)}{P\left(+/_{E}\right) * P(E) + P\left(+/_{noE}\right) * P(noE)} = \frac{0{,}95 * 0{,}02}{0{,}95 * 0{,}02 + 0{,}03 * 0{,}98}$$

$$= \frac{0{,}019}{0{,}019 + 0{,}029} = 0{,}396$$

Entonces, si a una persona estas pruebas le han dado resultado positivo —lo que en principio conduciría a diagnosticar esquizofrenia—, la probabilidad que efectivamente sea esquizofrénico es menos del 40%. Es posible que este resultado no esté cerca de la estimación intuitiva que uno haría y nos pone muy en alerta sobre la interpretación de pruebas de este tipo. El razonamiento intuitivo quizás nos habría llevado a creer que alguien a quien la prueba da positiva tiene muchas posibilidades de tener la enfermedad, pero no debemos confundir la probabilidad que la prueba de positiva si se tiene la enfermedad (*P(+/E)*) con la probabilidad de tener la enfermedad si la prueba da positiva (*P(E/+)*). Si una persona es esquizofrénica, la prueba le da positiva en un el 95% de las veces; pero si da positiva, la probabilidad que sea esquizofrénica es menor al 40%.
Este resultado no debe conducir a creer que la prueba no sirva para el diagnóstico. Por el contrario, ante una persona de la que no se tiene

[55] Nuevamente en este caso, esto puede conocerse a posteriori, luego de otras pruebas o del seguimiento del sujeto

ninguna información, la probabilidad que sea esquizofrénico es 0,02; cuando se agrega el dato que dice que el test dio positivo, la probabilidad que sea esquizofrénico asciende a 0,39. Nuevamente, se ve con claridad cómo este teorema permite usar los resultados de la experiencia para corregir probabilidades asignadas a priori.

En este ejemplo, la probabilidad de ser esquizofrénico para alguien que obtuvo resultado positivo en las pruebas es tan baja debido a la baja frecuencia de la esquizofrenia en la población general (prevalencia), pero demuestra lo equivocado que puede estarse si no se tienen en cuenta resultados falso positivo y falso negativo asociados a las pruebas diagnósticas.

Un abordaje alternativo a este problema es usando una tabla de doble entrada. Suponiendo que aplicamos el conjunto de pruebas a un universo de un millón de personas y usando las probabilidades enunciadas antes:

	Resultado de las pruebas		
	Positivo	Negativo	Total
Esquizofrénicos	19.000	1.000	20.000
No esquizofrénicos	29.400	950.600	980.000
Total	48.400	951.600	1.000.000

Queremos responder ¿cuál es la probabilidad que el sujeto sea esquizofrénico, si sabemos que la prueba le dio positiva? Para ello:

$$P\left(E/_{+}\right) = \frac{19.000}{48.400} = 0{,}39$$

Que es el mismo resultado que obtuvimos aplicando el teorema de Bayes.
La ventaja de la presentación a través de una tabla de doble entrada es que permite distinguir dos conjuntos de eventos sobre los que tenemos diferente conocimiento:
-Un *estado de realidad*, que es la condición de esquizofrénico o no esquizofrénico del sujeto. Este estado nos es desconocido.
-La evidencia observable, que está dada por el resultado de la prueba que aplicamos, que conocemos.
Como las pruebas nos son perfectas, los resultados deben leerse en términos probabilísticos y no determinísticos.

Cuando ingresemos a inferencia estadística veremos que ésta es la situación más frecuente: que dispongamos de cierta evidencia y debemos usarla para tomar una decisión acerca de un estado de realidad al que no conocemos.

Como puede verse, se trata de un teorema de gran importancia, por las consecuencias que tiene para muchas pruebas diagnósticas que se usan a menudo. Lo que hemos encontrado también implica el cuidado con que deben leerse los resultados de pruebas de cualquier tipo: diagnósticos, dosaje de productos prohibidos en deportistas, pruebas genéticas, etc. Para una correcta interpretación de los resultados de esas pruebas se deben conocer cuáles son los errores de tipo "falso positivo" y "falso negativo" que las acompañan.

Actividad práctica de repaso 6

1. En un estudio realizado por Belló y colaboradores (2005) se evaluó la prevalencia de depresión en la población mexicana. Los resultados obtenidos fueron los siguientes.

	f	f ´
Con Depresión	1741	0,045
Sin Depresión		0,955
Total	38700	

Complete los datos faltantes de la tabla y señale cuál es la probabilidad de que un mexicano padezca depresión.

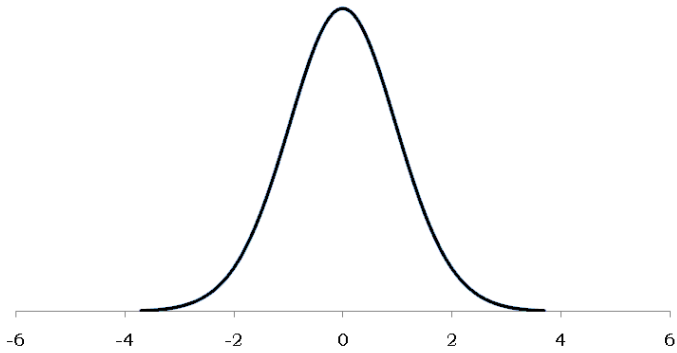
2. En un trabajo realizado por Icaza y colaboradores (2005) se examinó la prevalencia de diferentes tipos de suceso violentos. Algunos de los resultados obtenidos fueron los siguientes:

	Hombre		Mujeres	
Violación	14	0,01	78	0,08
Maltrato Familiar	204	0,17	227	0,23
Maltrato por pareja	9	0,01	133	0,14
Maltrato por otros	136	0,11	40	0,04
Secuestro	42	0,03	8	0,01
Accidente de tránsito	323	0,26	182	0,18
Enfermedad Grave	106	0,09	128	0,13
Sufrió un asalto con arma	393	0,32	188	0,19
Total	1227	1	984	1

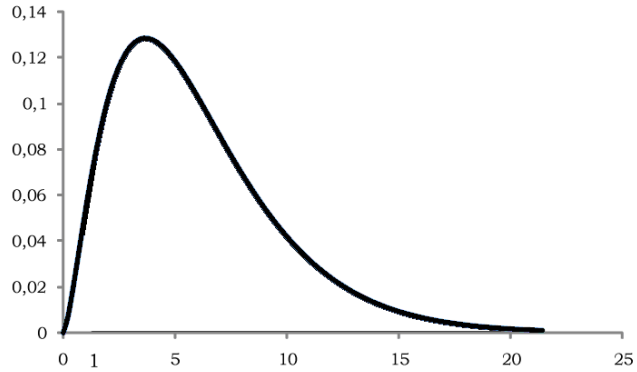
a. ¿Si elegimos al azar 100 mujeres cuántas habrían sufrido un maltrato familiar? ¿Y si elegimos a 100 hombres?
b. ¿Cuál es el suceso violento que tiene mayores probabilidades de ocurrir en los hombres? ¿Y en las mujeres?

3. Observando los siguientes gráficos indique a qué tipo de distribución están haciendo referencia. Comente alguna de sus características:

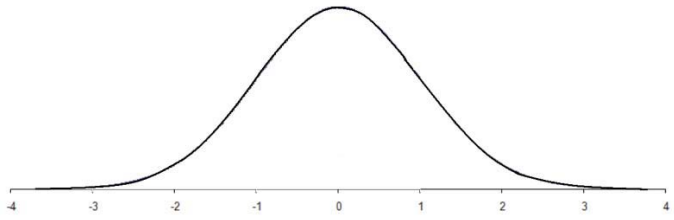
a.



b.

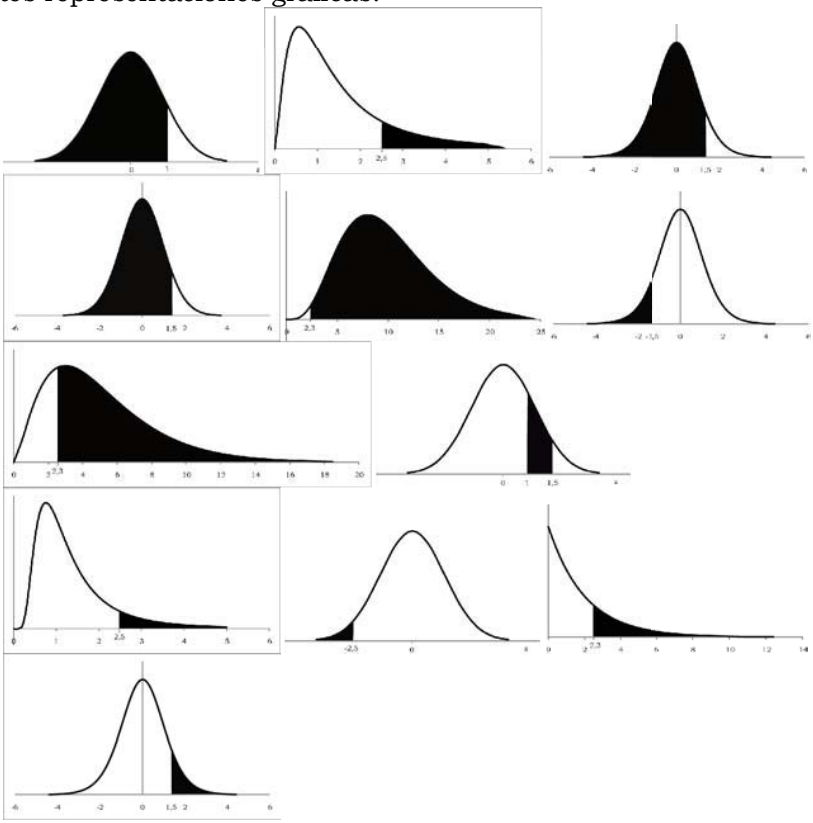


c.



4. Haga un esquema que represente el área bajo la curva normal que representa cada una de las siguientes probabilidades.
$P(z<2)$, $P(z>1,8)$, $P(-1<z<2)$, $P(z<-1)$, $P(t_5<1)$, $P(t_{30}<1)$, $P(t_{10}>2)$, $P(t_{10}<-2)$, $P(\chi^2_2>2)$, $P(\chi^2_{10}>2,3)$, $P(F_{5,7}>1)$

5. Identifique de manera aproximada las áreas indicadas en las siguientes representaciones gráficas:



Capítulo 7: Técnicas de muestreo

Waldino Romero
Eduardo Bologna

Nos aproximamos a la estadística inferencial, la que nos permitirá efectuar generalizaciones utilizando los datos que nos proporciona la estadística descriptiva. Mediante un conjunto de procedimientos, se podrán extender las conclusiones que se obtienen de un grupo reducido de casos al conjunto total que es objeto de estudio.
Es común en las investigaciones estudiar solamente un subconjunto de elementos y generalizar las conclusiones obtenidas al total; esto se debe a que estudiar uno o varios atributos o variables en todos los elementos requiere de fuertes inversiones e implica mayores tiempos, por ello es que se procede estudiando tales variables en una muestra y se estiman por inferencia los valores de la población. Se reducen así los costos y se disminuye el tiempo necesario para llevar adelante las investigaciones, si se respetan los procedimientos de la estadística inferencial el error que proviene de trabajar sólo con una parte de la población puede ser conocido y controlado.
El muestreo es un conjunto de procedimientos mediante los cuales se selecciona de un universo determinado, llamado población, un subconjunto que recibe el nombre de muestra, con el objetivo de llegar al conocimiento de determinadas características de los elementos de la población a través de la observación y generalización de esas características presentes en los elementos de la muestra.
Este recurso no es ajeno a la vida cotidiana. Aunque no se lo nombre de esta manera es lo que se realiza, por ejemplo, cuando un comprador examina sólo una parte de la mercadería que quiere llevar suponiendo que el resto posee las mismas características que las observadas. O bien, al afirmar que un pantalón de una marca "X" es bueno por tener una experiencia gratificante con una remera y un abrigo de la misma marca.
La primera idea que tenemos sobre el requisito que debemos pedir a una muestra para que las conclusiones sobre el total de la mercadería o la calidad del pantalón sean correctas es que es necesario que la muestra posea características similares a la población de la cual procede, precisaremos esta idea a lo largo de este capítulo.

Cuando la población es homogénea o sus elementos se encuentran bien mezclados no se requieren mayores cuidados al seleccionar una muestra dado que, en general, representará convenientemente a la misma. Este es el caso de una población conformada por elementos iguales entre sí en la característica de nuestro interés. Por ejemplo, en un frasco con 100 caramelos de frutilla, si extraemos 10 caramelos todos tendrán gusto a frutilla, por lo tanto la muestra es representativa. Por ser una población absolutamente homogénea, cualquier muestra será adecuada para conocer las características de toda la población (el gusto de los 100 caramelos). Si se trata de 50 caramelos de frutilla y 50 de ciruela, bien mezclados dentro del recipiente y extraemos 10, podemos creer que es muy probable que tengamos aproximadamente la mitad con gusto a manzana y la otra a ciruela.
Pero ¿qué ocurriría si en el frasco hay 50 caramelos de frutilla y 50 de ciruela y los de frutilla están todos cerca de la boca del frasco? Al extraer 10 caramelos es muy probable que concluyamos que todos tienen gusto a frutilla, dado que la población no está constituida por elementos bien mezclados y el muestreo no se ha realizado correctamente. Esto no ocurriría si extraemos del fondo del frasco 5 caramelos y 5 desde arriba, entonces esta vez la muestra sería representativa. En este ejemplo se aprecia que no podemos saber si la muestra representa bien o no a la población, porque deberíamos conocer la distribución de caramelos dentro del frasco, lo que solo sería posible si hubiésemos observado a toda la población (el frasco completo de caramelos). Se aprecia, sin embargo, cómo depende del proceso de muestreo y de la forma de la población, que los caramelos seleccionados sean representativos, conociendo así —al menos de manera aproximada— características de la población sin tener que trabajar con todos los elementos.

Los seres humanos tienen características muy dispares entre sí, es por ello que en las Ciencias Sociales es de sumo interés conocer las técnicas de muestreo para obtener muestras representativas de la población objeto de estudio.

En la obtención de una muestra, podemos identificar dos subprocesos. El primero es la selección, que consiste en operaciones mediante las cuales se incluyen algunos elementos de la población en la muestra. El segundo es la estimación, donde a partir de lo observado en la muestra estimamos los valores de la población, a ello se lo denomina estimación de parámetros y será tratado en los

próximos capítulos, limitándonos a la selección de la muestra en el presente.

Definiciones preliminares

Población

Utilizaremos la palabra **población** (o, indistintamente, universo) para designar, de manera genérica, a un conjunto de unidades de análisis que son objeto de un estudio particular. Tal conjunto puede estar definido con precisión en el tiempo y el espacio o no, a él se referirán los resultados obtenidos en la investigación por muestreo. Ejemplos de estos son: los pacientes con trastornos alimenticios del hospital Misericordia de la ciudad de Córdoba en el año 2009, o los votantes en las elecciones para intendente de la ciudad de Rio Cuarto en 2008, los ingresantes a la carrera de Medicina, los docentes de Argentina.

Las unidades de análisis antes mencionadas constituyen los elementos de la población, y pueden ser personas, hogares, instituciones, ciudades, etc.

Veamos a través de un ejemplo que es posible definir diferentes unidades de análisis. Sea el caso de una investigación acerca de la oferta educativa de nivel medio en Córdoba en 1998, la población estará constituida por todas las escuelas de ese nivel en la provincia, mientras que si el estudio se dirige a las características de los docentes de ese nivel, la población será la de los profesores de nivel medio que ejercen en la provincia de Córdoba. Por último, dentro del mismo ejemplo, si el objetivo de la investigación es el de analizar el nivel socioeconómico de los alumnos, la población se conformará por los hogares de los que provienen los estudiantes de nivel medio de la provincia de Córdoba.

Veamos ahora que el tamaño que tiene una población es un factor muy importante para una investigación, dependiendo de la cantidad de elementos que posea la misma puede ser tratada como finita o infinita. Se dice que es finita si el número de elementos es limitado y se puede tener acceso a todos o casi todos los elementos, ejemplo de ello serían los alumnos de cuarto año del colegio Manuel Belgrano del ciclo lectivo 2000 o egresados de la carrera de ingeniería electrónica de la U.N.C. en el año 2007. Se habla de población infinita[56] cuando el número de elementos que integra la misma es elevado, por ejemplo, los automovilistas que circularon por las rutas nacionales en el 2008,

[56] La idea de infinito no es la de la matemática, sino que se refiere a una relación pequeña entre la cantidad de casos de muestra y los de la de la población

o alumnos de nivel inicial de la República Argentina. En ciertas ocasiones no se especifica con precisión el tiempo por lo que se habla de población hipotética, tal caso sería el de alumnos de cuarto año del colegio Manuel Belgrano (sin especificar ciclo lectivo), por lo que no sólo se incluyen los elementos existentes hoy sino también los que en el futuro lo hagan. En este caso, la población puede siquiera enumerarse. Así sucede también con estudios que plantean probar los efectos de un tipo de psicoterapia sobre pacientes diagnosticados como esquizofrénicos. En ese caso la "población" es la de las personas diagnosticadas como esquizofrénicas en la actualidad y también aquellas que lo serán en el futuro. En casos como estos no tenemos posibilidad de delimitar la población completa, decimos que se trata de **poblaciones hipotéticas**.

Para analizar características de las unidades de análisis, puede creerse que lo mejor sería realizar un **relevamiento exhaustivo** de las mismas. Este relevamiento consistiría en observar dicha característica (variable) en cada uno de los individuos de la población. Tal modo de recolectar la información se conoce con el nombre de **censo**.

Desde el punto de vista práctico, es frecuente realizar relevamientos exhaustivos cuando las poblaciones son de pequeño tamaño y están bien delimitadas en el tiempo y en el espacio. Un ejemplo de este caso sería la situación en que se pretende conocer la opinión de los alumnos de un curso acerca de su docente, aquí lo más adecuado sería indagar (vía encuesta) a cada uno de ellos.

Sin embargo, la mayoría de las poblaciones de interés para la investigación social tienen dimensiones considerables (todos los alumnos de primer grado del país) o son inaccesibles (los consumidores de una marca de gaseosa) o son hipotéticas (las personas diagnosticadas de depresión).

Cuando se busca dar generalidad a los resultados alcanzados a través de experimentos, la población de referencia suele ser hipotética: los efectos de una droga sobre la depresión, la efectividad de un tipo de psicoterapia dirigida a pacientes diagnosticados como esquizofrénicos. Son casos que aspiran a generar resultados válidos para una población hipotética, por lo que el relevamiento completo es imposible y solo se puede observar a unos pocos casos. El modo en que seleccionemos esos pocos casos es de la máxima importancia para que los resultados sean válidos para los no observados también.

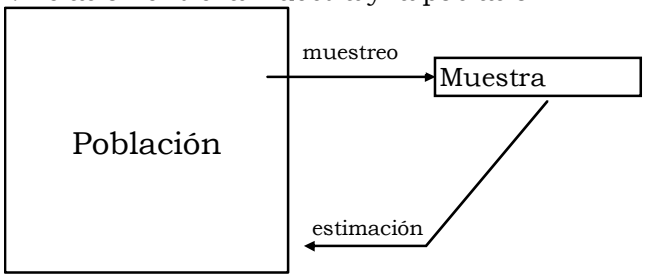
Además de permitir economía de recursos, el muestreo permite profundizar los estudios que se realizan sobre los elementos de la muestra. En efecto, en los relevamientos exhaustivos de gran magnitud, como los censos nacionales o provinciales, no es posible profundizar con la indagación; por el contrario se deben hacer pocas

preguntas a fin de alcanzar a una gran población en un tiempo aceptable. Si se trabaja con una muestra, se puede indagar con más detalle y profundidad, y luego, con los procedimientos adecuados, extender los hallazgos a la población de la que la muestra proviene.

Muestra

Se llama **muestra** a un subconjunto de una población que comparte sus características en los aspectos de interés para la investigación. El concepto de muestra va ligado al de representatividad, es decir a su capacidad de actuar como "representante" de los elementos de la población que no han sido seleccionados. Tal representatividad no implica una identidad en todos los aspectos, son solamente aquellas características que se encuentran bajo análisis las que deben ser compartidas por la muestra y la población.

Esquema 1: Relación entre la muestra y la población



Para fundamentar la necesidad del muestreo y del cuidado que debe ponerse en la construcción de la muestra, presentaremos brevemente algunos principios sobre estimación, que serán retomados y tratados con más detalle en el capítulo 8.

El objetivo de una investigación por muestreo es obtener información acerca de una característica de la población a partir de datos provenientes de la muestra. La característica poblacional que pretende conocerse se llama **parámetro**.

Como se señaló, a través del proceso de muestreo se seleccionan algunas unidades del universo que actuarán como representantes de la población completa. Esta muestra será toda la información disponible para realizar estimaciones acerca de los parámetros poblacionales.

Los valores calculados sobre los datos muestrales son los que se utilizarán para realizar aproximaciones a los valores poblacionales; tales valores son denominados **estimadores puntuales**, o simplemente **estimadores.**

De una población se pueden obtener un número muy grande de muestras, dependiendo de su tamaño y del universo de donde se extraen, pero en un proceso de muestreo se toma una sola muestra de todas las posibles.
Veamos esto a través de un ejemplo: consideremos una población constituida por sólo tres elementos, 3 personas (Juan, Pedro y Marcos). De esta población se extraerán muestras de dos elementos (personas) cada una. Como vimos en el capítulo 6, los elementos de cada muestra pueden extraerse con o sin reposición, en el próximo capítulo usaremos el primero de ellos para ilustrar la forma en que se distribuyen las muestras. Sin embargo, en las aplicaciones prácticas el muestreo se realiza sin reposición ya que repetir uno o varios elementos atenta contra la heterogeneidad de la muestra. En nuestro ejemplo, el muestreo sin reposición da lugar a las combinaciones: Juan-Pedro, Juan-Marcos y Pedro-Marcos. Así, sólo hay tres muestras posibles de tamaño dos, trabajando sin reposición.
Dado que los datos muestrales se utilizarán para estimar los valores poblacionales, cada una de las muestras proveerá de una estimación diferente. Para un parámetro elegido, habrá tantos estimadores como muestras puedan extraerse de la población (con algunos resultados muestrales repetidos). Esta variación (casi) impredecible de los estimadores, hace que se comporten como variables aleatorias.
Es posible (aunque esperamos que sea poco probable) que el estimador se aleje notablemente del valor poblacional al que se pretende estimar. Así, si se busca estimar la edad promedio de los estudiantes de cierta carrera universitaria y no se cuenta con un registro de ellos, puede usarse una estimación a partir de una muestra extraída de todos los estudiantes. El promedio calculado sobre esos datos se tomará como estimación de la edad de todos los estudiantes; pero puede ocurrir que la muestra contenga, por azar, al grupo de mayor edad de toda la población, con lo cual el resultado así obtenido será una sobreestimación del verdadero valor poblacional. Debe destacarse que este hecho resulta completamente inadvertido para el investigador. Con respecto a la ocurrencia de estos eventos que conducen a resultados engañosos, sólo será posible —bajo ciertas condiciones—calcular sus probabilidades.
Dos aspectos importantes para recordar cuando se usan muestras:
La variabilidad de las muestras: Las muestras difieren entre sí. Varias muestras extraídas de la misma población pueden conducir a resultados diferentes.

La representatividad de la muestra: No indica la similitud entre la muestra y la población, sino su obtención por un procedimiento

adecuado que permita estimar valores de la población a partir de lo que se halle en la muestra. Por ahora es suficiente recordar que la calidad de la muestra mejora cuando incluye a más observaciones.

Características de los estimadores
Si fuera posible extraer todas las muestras de una población, se contaría con todos los estimadores de un parámetro dado. Cuando el promedio de todos los valores obtenidos en todas las muestras de un determinado tamaño para cierto parámetro es igual al valor de esa característica en la población, se dice que el estimador muestral es **insesgado**. Se denomina sesgo de un estimador a la diferencia entre el promedio que alcanzaría sobre todas las muestras posibles y el verdadero valor del parámetro poblacional. En el caso de un estimador insesgado esta diferencia es igual a cero (sin sesgo). Profundizaremos este concepto el próximo capítulo.
Intuitivamente puede pensarse que el aumento en el tamaño de la muestra (número de elementos que la componen) mejora la calidad de la estimación. Esta característica, que se cumple para algunos estimadores, es denominada **consistencia**. Un aumento en el tamaño de la muestra no garantiza que se obtendrán resultados más próximos a los valores poblacionales: por grande que sea la muestra no resulta imposible que a ella pertenezcan los valores más extremos de la población. Veamos esto con un ejemplo. Supongamos que se quiere estimar la calificación promedio con que egresan los alumnos de una carrera dada y que se toma una muestra de 30 egresados. Puede ocurrir que la muestra (por azar) contenga a los treinta mejores promedios de la carrera, con lo que se obtendría una sobreestimación del verdadero promedio de toda la población. El concepto de consistencia expresa que si la muestra, en lugar de 30 elementos, contiene 100, la probabilidad de obtener una sobreestimación igualmente extrema es menor. Intuitivamente podemos ver que resulta más factible encontrar a los 30 mejores que a los 100 mejores.
No debe entenderse que el estimador alcance valores más próximos a los paramétricos a medida que la muestra aumenta de tamaño, sino que disminuye la probabilidad de obtener resultados muy lejos de los paramétricos a medida que la muestra crece. Esta idea parece un poco abstracta ahora, se aclarará cuando la usemos en casos aplicados.

Cuando se planifica la extracción de una muestra, no se conoce en detalle a la población, por lo que no es posible saber a priori cómo debe ser la muestra que la represente. Si hay algunas características de la población que sean conocidas, entonces las reproduciremos en

la muestra. Por ejemplo, si sabemos que en la población de docentes primarios de una ciudad, hay 80% de mujeres, entonces, cuando construyamos una muestra ése debe ser el porcentaje de mujeres en ella. Pero ¿Cuántas maestras de menos de 30 años debemos incluir en la muestra? No lo sabemos, porque no conocemos esa característica de la población. Sobre estas características de la población –las que no conocemos—la mejor garantía de representatividad es seleccionar a los individuos a través del azar.

La posibilidad de extrapolar los resultados muestrales a la población completa está así ligada a la posibilidad de asignar probabilidades a cada una de las muestras que podrían seleccionarse de una población. De manera que sólo será lícito utilizar los resultados muestrales como estimación de valores poblacionales cuando sea posible conocer a priori cuál es la probabilidad que cada individuo de la población tiene de ser incluido en la muestra. Las técnicas de muestreo que permiten asegurar este requisito se denominan de carácter **probabilístico**. Otras técnicas que no cumplen esta condición se llaman muestreos **no probabilísticos** y muchas veces son utilizadas con el objetivo de reducir tiempos y costos, pero los resultados obtenidos de ellas no pueden utilizarse para extraer conclusiones acerca de la población. Son adecuados en estudios que no buscan generalidad de sus conclusiones sino profundidad en el análisis de los casos observados. Sin embargo, hay investigaciones que usan muestras no probabilísticas para dar carácter general a los resultados que se obtienen. Cuando se hace esto, nunca hay certeza del grado de generalidad de lo que se concluye y hasta qué punto no se trata de características específicas de los individuos a los que se observó, porque no puede conocerse el margen de error de lo que se estima. Cuando las limitaciones prácticas hacen imposible la obtención de una muestra probabilística, se deben tomar recaudos para que la muestra sea heterogénea y que la inclusión de casos dependa lo menos posible de la voluntad de los participantes.

Para que sea posible asignar probabilidades a cada una de los individuos de formar parte de la muestra, el método por el cual los individuos son seleccionados debe excluir la elección voluntaria. Esto implica que el proceso de elección debe quedar estrictamente librado al azar, sin dejar margen para que se filtre la intencionalidad, no se trata de elegir a “cualquiera”. En consecuencia, caerán en la categoría de técnicas de muestreo no probabilísticas todos los procedimientos en que exista alguien que pueda tomar la decisión acerca de si una unidad va a ser o no incluida en una muestra.

Muestreos probabilísticos: las muestras obtenidas por estos procedimientos permiten generalizar los resultados obtenidos en ellas a toda la población de referencia. El requisito para que una muestra sea probabilística es que sus elementos hayan sido elegidos al azar (aleatoriamente), sin la participación voluntaria que decida a quién incluir y a quién excluir de la muestra.
Muestreos no probabilísticos: en estas muestras no se cumple el requisito de aleatoriedad en la selección de los elementos que la componen. Los resultados no se pueden generalizar de manera probabilística más allá de los casos observados.

Muestreos probabilísticos

Cuando se pretende que los resultados obtenidos en una muestra puedan ser generalizados a la población de la cual ésta proviene, se hace necesario recurrir a muestreos de tipo probabilísticos. Se trata de diseños que requieren más cuidado en su elaboración y de manera concomitante son más costosos. Señalamos a continuación los tipos puros de muestreos probabilísticos y luego veremos algunas combinaciones de ellos que se usan a menudo.

Muestreo irrestricto aleatorio o Aleatorio simple
Se trata de una técnica que asigna igual probabilidad de pertenecer a la muestra a todos los individuos de la población. Para su realización se requiere contar con una lista de los elementos de la población. Este listado es lo que se denomina el **marco de la muestra**. Dicho marco debe cumplir con los requisitos de exhaustividad (es decir, que sea completo) y debe cuidarse que no existan duplicaciones. La tarea de depuración de un listado puede insumir bastante tiempo según el tamaño de la población y el tipo de marco muestral de que se trate. Es diferente el caso de un registro electoral del total de historias clínicas de un gran hospital o del listado de todos los alumnos de una facultad).
Su realización consiste en numerar los elementos del listado y elegir aleatoriamente una cantidad n de ellos (el tamaño de la muestra). La aleatoriedad de la elección queda garantizada por el uso de una tabla de números aleatorios o de un generador electrónico de los mismos, la mayoría de las calculadoras de bolsillo disponen de esta opción en la tecla #RAN (por random, aleatorio).
La exigencia de contar con el listado de los elementos de la población es una limitación importante para este tipo de muestreo por lo que, como se verá más adelante, a menudo se lo utiliza en combinación con otros procedimientos. Además, puede ocurrir que los elementos

elegidos en la muestra se encuentren dispersos geográficamente de manera que resulte notablemente costoso ubicar a cada uno de ellos para entrevistarlos.
Una ventaja importante es que no exige que se conozca a priori ninguna característica de la población. Desde el punto de vista estrictamente estadístico, resulta más fácil acotar el error de muestreo ya que las distribuciones de probabilidad subyacentes son bien conocidas para este tipo de muestreo.

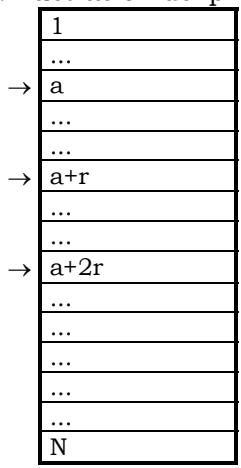
Muestreo sistemático
Un procedimiento alternativo al muestreo irrestricto aleatorio lo constituye el muestreo sistemático. En este caso sólo se selecciona un elemento aleatoriamente y, comenzando por él, se recorre el marco de la muestra tomando los elementos siguientes a intervalos regulares. El primer paso consiste en determinar el número de veces que puede incluirse la muestra en la población. Este valor es el que resulta de dividir el tamaño de la población en el tamaño de la muestra:

$$r = \frac{N}{n}$$

Donde:
r representa a las veces que la población contiene a la muestra,
N es el tamaño de la población,
n es el tamaño de la muestra.

Luego se genera un número aleatorio entre 1 y r, éste constituye el primer elemento de la muestra al que llamaremos "a". A continuación se selecciona el elemento que se encuentra r unidades más adelante en la lista (el elemento que ocupa el lugar $a+r$) y así sucesivamente hasta recorrer la lista completa, lo cual ocurrirá cuando se haya alcanzado al enésimo y último elemento de la muestra.
El gráfico siguiente ilustra este procedimiento, las flechas a la derecha señalan los primeros elementos que constituyen la muestra; el procedimiento prosigue de idéntica manera hasta la última unidad de la muestra.

Esquema 2: Ilustración del procedimiento de muestreo sistemático



Esta técnica tiene, en principio, los mismos requerimientos que el muestreo irrestricto aleatorio, en cuanto a la necesidad de disponer del marco de la muestra, pero puede adaptarse cuando no se cuenta con la lista.

Su uso se justifica por dos razones. La primera es que si la lista no presenta ninguna tendencia especial, el muestreo sistemático facilita la extracción de la muestra, simplifica la operación, ya que la elección de todos los números aleatoriamente puede ser lenta por la aparición de elementos repetidos que deben descartarse. La segunda razón es que puede ocurrir que el listado este ordenado según algún criterio (por fechas, o por edades de los individuos, etc.), de ser así, existe el riesgo que una muestra irrestricta aleatoria concentre los elementos elegidos en alguna "zona" de la lista, sobrerepresentando de esta manera a los individuos que tienen alguna característica en común. En este caso el muestreo sistemático asegura que el marco de la muestra sea recorrido completamente a intervalos iguales.

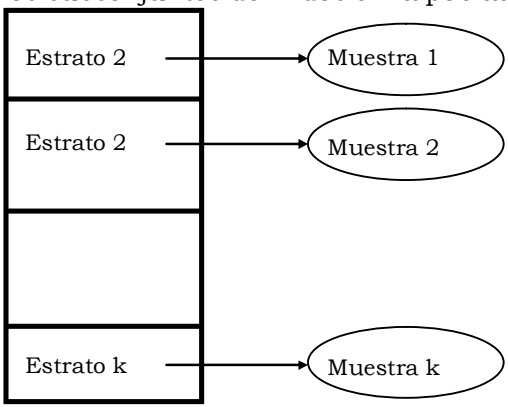
Por otra parte, existen situaciones en que el muestreo sistemático es desaconsejado: diremos que la población presenta un comportamiento cíclico si cierta característica se repite regularmente cada cierta cantidad fija de casos. Un ejemplo de esto es una muestra a lo largo del tiempo, en que algunos "momentos", tales como inicios o finales de mes, determinados días de la semana, etc., se repiten periódicamente en el año. Sea que tratamos de estimar los ingresos a un hospital y elegimos, de manera sistemática, los días que constituirán la muestra. Si el valor de r fuera 7 ó un múltiplo de él, la muestra contendría siempre al mismo día de la semana y sería un sesgo grave, porque los

ingresos al hospital pueden ser muy diferentes en los diferentes días de la semana. Aún cuando el ejemplo va más allá de lo realista, su utilidad está en alertar contra el uso de procedimientos sistemáticos cuando se sospecha de la existencia de una tendencia cíclica en el listado con que se cuenta.

Por último, hay casos en que el muestreo sistemático se usa para prescindir del marco de la muestra. Es cuando se seleccionan casos a partir de poblaciones en movimiento: la cola para entrar al cine es un caso. Otro caso frecuente son las encuestas "boca de urna" en las que se pregunta por quién votó a personas que salen del lugar de votación. Allí no hay listado al que recurrir, es a la gente que está físicamente presente que debemos encuestar. El recurso consiste en elegir a una persona cada r, buscando que la muestra incluya a personas que llegaron temprano y que llegaron tarde, que votaron a la mañana y a la tarde, es decir que recorramos la "fila" alcanzando a quienes están en los primeros lugares y también a quienes están al final.

Muestreo estratificado
El procedimiento consiste en extraer muestras de subconjuntos de la población llamados estratos. Se espera que tales estratos sean homogéneos en su interior con respecto a alguna característica conocida a priori. Dicha característica se denomina criterio de estratificación.

Esquema 3: En el muestreo estratificado se eligen muestras diferentes de cada uno de los subconjuntos definidos en la población.



Un ejemplo puede aclarar esta idea: supongamos que se requiere estimar el consumo de bebidas alcohólicas de los jóvenes. Como se

supone que esa variable está influida por el nivel socioeconómico, puede utilizarse este último como criterio para estratificar a la población, dividiéndola en subconjuntos en los que el nivel socioeconómico sea homogéneo. La cantidad de estratos que puedan definirse de esta manera dependerá del grado de precisión con que pueda medirse el nivel socioeconómico. Si estratificamos con categorías nivel socioeconómico alto, medio y bajo, entonces será k=3 y seleccionaremos jóvenes de cada uno de los tres estratos.
Es de destacar que la división de la población en estratos demanda cierta información acerca de ella (en este ejemplo el nivel socioeconómico). Cuanta más precisión se pretenda lograr con la estratificación, tanto mayor será la información necesaria a priori.

A veces los estratos están dados de manera más inmediata, como sucede con los alumnos de una carrera universitaria. Si se busca una muestra que represente a los estudiantes de la carrera, es conveniente estratificar por año de cursado, de modo de incluir en ella alumnos de todos los años.
Este procedimiento de muestreo se utiliza cuando se busca aumentar la precisión de la estimación sobre la población total o bien para mejorar la precisión sobre los estratos individuales. Se obtendrá tanto mayor ganancia en la estimación cuanto más homogéneos sean los estratos en su interior y más diferentes (heterogéneos) sean entre ellos. Veremos en detalle las razones de esto cuando presentemos el error en la estimación por intervalo.

a. Muestreo estratificado con afijación igual o uniforme
Se llama afijación a la modalidad que se utilice para distribuir la muestra sobre los estratos definidos. Según la información con que se cuente a priori, se podrán utilizar diferentes formas de afijar, la primera que tratamos es la afijación igual o uniforme.

Ésta se aplica cuando no existe información alguna acerca de los estratos y no hay razón para ponderar especialmente alguno de ellos. El procedimiento es simplemente extraer la misma cantidad de casos de cada estrato. Las muestras tienen así el mismo tamaño que resulta:

$$n_i = \frac{n}{k}$$

Donde:
n_i es el tamaño de las muestras extraídas de cada estrato.
n es el tamaño de la muestra total.
k representa el número de estratos en que fue dividida la población.

En el ejemplo sobre el consumo de bebidas alcohólicas, si se trata de una muestra de 600 jóvenes, con este tipo de afijación, y como son tres estratos, extraeremos 200 casos de cada estrato.

b. Muestreo estratificado con afijación proporcional
Consiste en extraer de cada estrato una muestra cuyo tamaño resulte proporcional al estrato del que proviene. Para ello, primero calculamos f que es la proporción de población que integra la muestra:

$$f = \frac{n}{N}$$

Este cociente se llama **fracción de muestreo**, en su cálculo:
n es el tamaño de la muestra total.
N representa el tamaño de la población.

Una vez que conocemos la fracción de muestreo, la aplicamos a cada uno de los estratos para obtener la cantidad de casos que deben extraerse de cada uno:

$$n_i = f * N_i$$

Donde:
N_i es la cantidad de casos en el estrato i-ésimo (en la población)
n_i es la cantidad de casos que se extraerán del estrato i-ésimo
f es la fracción de muestreo que calculamos antes.

Para usar este tipo de afijación en el ejemplo del consumo de bebidas alcohólicas necesitaríamos conocer la cantidad de jóvenes que hay en cada estrato, es decir cuántos jóvenes de nivel socioeconómico alto, medio y bajo hay en la población. Este dato usualmente es desconocido.

Veamos un ejemplo con datos reales: en el año 2008 en Psicología, la siguiente era la distribución de alumnos según el año en que habían ingresado a la carrera (considerando solo a los ingresados a partir del año 2000).

Año de ingreso a la carrera	Cantidad de alumnos
2008	1.357
2007	1.113
2006	1.001
2005	807
2004	859
2003	764
2002	700
2001	483
2000	399
total	7.483

Fuente: Anuario estadístico 2008, UNC

Si queremos una muestra de 200 alumnos de esa población, la fracción de muestreo será: $f = \frac{200}{7483} = 0{,}027$ que redondeamos a 0,03. Aplicaremos esta fracción a cada estrato, es decir multiplicamos por esa fracción a la cantidad de casos que hay en cada estrato y redondeamos al entero más próximo.

Año de ingreso a la carrera	Cantidad de alumnos	Casos a seleccionar
2008	1.357	41
2007	1.113	33
2006	1.001	30
2005	807	24
2004	859	26
2003	764	23
2002	700	21
2001	483	14
2000	399	12
total	7.483	

Entonces, para la muestra debemos elegir 41 alumnos que hayan ingresado en 2008, luego 33 del 2007 y así sucesivamente. A esa selección la hacemos aleatoria, ya que es posible contar con el marco de la muestra que es el registro de alumnos que tiene la facultad.

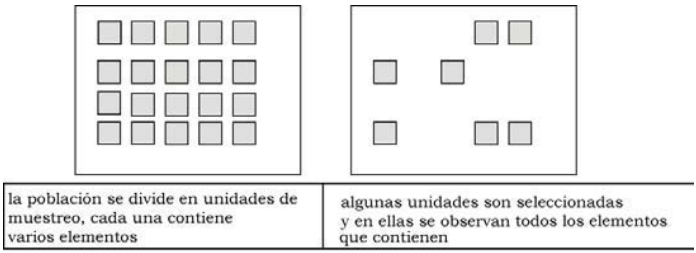
Existen otras formas de afijación, más complejas, que tienen en cuenta más información acerca de la población. Por ejemplo, la afijación óptima tiene en cuenta la heterogeneidad de cada estrato

para elegir la cantidad de casos que se extraen de cada uno. Recordemos el ejemplo de los 100 caramelos de frutilla dentro del frasco, cuando los elementos son muy homogéneos, bastan unos pocos casos para conocerlos a todos, por el contrario, cuanto más dispares (heterogéneos) sean los elementos de la población, tantas más observaciones serán necesarias para hacer estimaciones de buena calidad. Este es el criterio que se usa en el procedimiento llamado afijación óptima. Si se dispone de suficiente información, es posible incluir el costo de acceso a los casos pertenecientes a diferentes estratos y realizar una afijación que minimice el costo global del muestreo, para más detalles sugerimos Scheaffer et al (1987).
Solo nos ocuparemos de las dos formas de afijación que hemos mencionado: igual y proporcional.

Dos puntos que deben recordarse para decidir el uso del muestreo estratificado:
- Cada criterio de estratificación demanda información acerca de la población. Esa información debe provenir de un conocimiento anterior al muestreo.
- El muestreo estratificado implica un aumento de la calidad de la estimación bajo la condición que los elementos que están en un mismo estrato sean homogéneos entre sí y que los estratos sean diferentes unos de otros (heterogéneos entre sí).

Muestreo por conglomerados
Se denomina conglomerado a una unidad de muestreo que está constituida, en su interior, por varios elementos de la población. Para un diseño por conglomerado deben definirse unidades primarias de muestreo que contengan en su interior a las unidades elementales.
Un ejemplo clásico de este tipo de muestreo lo constituye el caso en que se requiere extraer una muestra de hogares en una ciudad. Se trata de una situación en la que resulta imposible contar con un listado de las unidades primarias (los hogares). Un muestreo por conglomerados permite resolver este problema tomando como unidad de selección a las manzanas urbanas. Numerando estas unidades sobre un plano de la ciudad (o del área que se pretenda relevar) se selecciona aleatoriamente un número *n* de ellas. Una vez identificadas las unidades primarias que constituirán la muestra (las manzanas) se relevan todos los hogares (unidades elementales) que residen en cada una de ellas.

Esquema 4: Principio del muestreo por conglomerados



Otro ejemplo es aquel en el que necesitamos una muestra de 100 alumnos de primer grado de la provincia de Córdoba. Tenemos dos problemas: el primero es que no hay una lista de ellos, no contamos con un marco muestral. El segundo es que, aunque la tuviéramos, tendríamos que ir a buscar a cada uno de los 100 alumnos muestreados a su ciudad, dentro de la provincia, lo que tendría un costo enorme. En este caso, el conglomerado (o unidad primaria) que se elije es la escuela: en lugar de elegir 100 alumnos al azar, elegimos 10 escuelas al azar del listado de ellas, que sí existe. Luego, en cada escuela elegimos 10 alumnos, también al azar, a partir del listado de alumnos de primer grado, que sí está disponible en cada escuela.

Los requisitos para la efectividad de este tipo de muestreo pueden ser resumidos como sigue:
- Las unidades primarias de muestreo deben ser homogéneas entre sí, de tal manera que cada unidad sea intercambiable con otra.
- Cada unidad primaria debe contener, en su interior, la heterogeneidad propia de la población que se releva.

Volviendo al ejemplo anterior en que las unidades primarias son las manzanas, se espera que cada manzana sea similar a las vecinas en cuanto a su composición (dentro de un área geográfica dada, se trata de un supuesto válido). Además, cada manzana dispondrá, en su interior, de la variedad que sea propia del área en cuestión (en términos de su composición: número de hogares particulares, comercios, escuelas, etc.).
Además del ejemplo de las manzanas, otros conglomerados utilizados son: establecimientos educativos, instituciones de cualquier tipo, etc. En general, todo conjunto que contenga en su interior a las unidades elementales puede elegirse como conglomerado.

Uso combinado de técnicas de muestreo

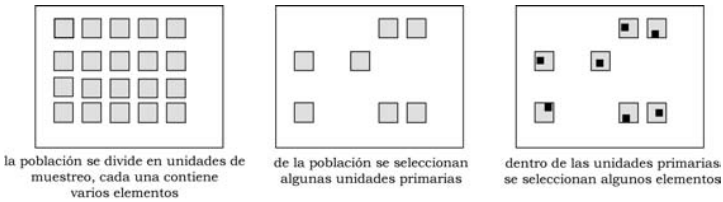
Las modalidades de muestreo presentadas pueden combinarse de varias formas para resolver problemas particulares.

En el muestreo por conglomerados puede usarse una segunda etapa posterior a la selección de las unidades primarias consistente en la incorporación en la muestra sólo de algunas de las unidades elementales contenidas en ellas. Tal elección puede realizarse por un diseño aleatorio irrestricto o sistemático.

Aplicando este concepto al ejemplo anterior, la situación consistirá en la elección de algunos hogares dentro de las manzanas seleccionadas (en lugar de relevarlos a todos). Para lograr esta muestra, se instruye a los encuestadores para que, comenzando por una esquina dada, entrevisten a un hogar cada una cantidad dada (uno cada diez por ejemplo para contar con una muestra del 10% de los hogares de la manzana).

O bien, en el ejemplo de los alumnos de primer grado, se trataría de seleccionar a 10 de ellos en cada escuela de manera aleatoria.

Esquema 5: Muestreo por conglomerados en dos etapas



La obtención de una muestra en una situación real implica la combinación de varios de estos procedimientos. Por ejemplo, para un sondeo de opinión sobre la población adulta de una ciudad, el primer paso es estratificar a la población (los adultos residentes en la ciudad), según nivel socioeconómico, porque esta es una variable que suele hacer diferencias marcadas en opinión política o en pautas de consumo. Un dato que puede tenerse sobre el nivel socioeconómico de los habitantes, es el provisto por el censo de población, que incluye preguntas que permiten hacer una caracterización de los hogares según esta variable. Sin embargo, el censo se hace cada diez años y se desactualiza al poco tiempo. Por eso suele usarse, como aproximación al nivel socioeconómico, el área de residencia. Si bien no es absoluta, la división puede aproximarse con barrios habitados mayoritariamente por hogares con nivel socioeconómico más alto y más bajo. Así, en un primer momento es posible dividir la ciudad en estratos, determinados por grandes áreas que aproximan el nivel socioeconómico. Como es posible conocer la cantidad de población

que vive en cada una de esas áreas, el muestreo estratificado se puede hacer con estratificación proporcional.
En la etapa siguiente, en lugar de seleccionar a las personas que responderán a la encuesta, seleccionamos las manzanas. Las enumeramos dentro de cada estrato y hacemos una selección aleatoria de la cantidad de ellas que sea necesaria para nuestro tamaño de muestra. Una vez seleccionadas las manzanas, a cada encuestador se le indica que recorra la que le tocó en un sentido establecido de antemano y que llame a la puerta en un domicilio cada 10, por ejemplo. En cada domicilio seleccionado se debe preguntar por las personas adultas que viven allí y se debe seleccionar aleatoriamente a una de ellas, quien será finalmente la que responda a nuestro cuestionario.
A este esquema general deben agregarse precisiones para casos particulares, como el rechazo a responder o la falta de gente en el domicilio seleccionado. Lo usual es volver dos veces luego de un rechazo y si se fracasa las dos veces, entonces ese domicilio se reemplaza por otro, según una regla establecida de antemano. También hay que aclarar desde el principio qué hacer en los casos de edificios, barrios cerrados, etc.
Para los sondeos de opinión, en una ciudad mediana, se estila usar 400 casos en 20 conglomerados de 20 casos cada uno.
Como se ve, lo que se busca es dejar el menor margen posible para improvisar, para evitar que la selección sea voluntaria. Esto se hace para asegurar que la muestra sea aleatoria y por tanto probabilística.
Si se busca una muestra de pacientes internados en centros de salud de una determinada región, se empezará por el listado de esos centros, a los cuales es posible estratificar según su dependencia pública o privada, según su tamaño, o según otra característica de interés. Una vez establecida la estratificación, se seleccionan al azar una cierta cantidad de centros de salud de cada estrato y, en cada centro será necesario contar con el registro de pacientes internados, aquí se introducirán las restricciones propias de cada investigación, el período durante el cual nos interesan los datos, las patologías que fueron motivo de la internación, etc. La selección de casos (historias clínicas de los pacientes internados) puede realizarse de manera irrestricta aleatoria si se dispone del listado.

Muestreos no probabilísticos

Los procedimientos que pretenden estimar valores de la población a partir de resultados muestrales sólo pueden aplicarse cuando es posible asignar probabilidades a priori a los individuos de ser parte de

la muestra. Aún cuando no cumplen con este requisito, las técnicas de muestreo no probabilísticas se utilizan en situaciones en que se pretende adquirir un conocimiento inicial de un problema que se encuentra escasamente delimitado, como sucede en los estudios de tipo exploratorio. También es frecuente su uso cuando no se pretende un conocimiento acabado sobre la población completa sino que el objetivo es profundizar sobre algunos casos en particular (estudios de casos) a través de entrevistas en profundidad que reconstruyan la historia de vida de algunas personas solamente. Un tercer uso es el de aproximar los resultados que se obtendrían con muestras probabilísticas. Como señalamos antes, deben tenerse recaudos con la generalidad de las conclusiones provenientes de este tipo de muestreo.

Una clasificación útil para este tipo de muestreo toma en consideración su definición, esto es que la elección no depende del azar sino de la voluntad de alguien. De esta manera los muestreos no probabilísticos pueden clasificarse según quien sea el responsable de la elección en: muestreo por cuotas, de juicio o intencional, autoelegido y accidental.

Muestreo por cuotas

En este muestreo se busca reproducir de la manera lo más ajustada posible las características de la población en la muestra. Se llama cuotas a las fracciones de la muestra con las distintas características. Este muestreo se usa cuando se conocen algunas características de la población y se busca reproducirlas en la muestra. Por ejemplo, si la muestra contiene una proporción de varones y mujeres igual a la que hay en la población y además respeta las proporciones de diferentes grupos de edad, decimos que usamos cuotas de sexo y edad. A estos criterios suele agregarse otros, según los objetivos de la investigación, por ejemplo, en las encuestas políticas, es frecuente establecer cuotas de acuerdo al voto en las elecciones pasadas. Los ejemplos que indicamos antes, como reproducir la proporción de varones y mujeres entre los docentes primarios, o de las diferentes proporciones de niveles educativos de la población en la muestra, son casos de muestreo por cuotas.

En la práctica, se encarga a los encuestadores una cantidad especificada de casos a relevar, sin indicar quiénes serán esos casos ni cómo hallarlos, esto le da el carácter de no probabilístico, pero respetando las cantidades de casos en cada cuota. El entrevistador elige a “cualquiera”, según su comodidad o preferencias, es decir que no es posible conocer a priori qué probabilidad tienen los elementos

de la población de caer en la muestra, pero esa elección está limitada por la exigencia que los elementos reúnan ciertas características.
Un ejemplo es en el que se indica a un encuestador que complete 10 entrevistas a médicos que trabajen en relación de dependencia en la ciudad de Córdoba que tengan hasta diez años de egresados; a otro, que entreviste a 15 médicos también en relación de dependencia, pero que tengan más de diez años de egresados. Si bien en este caso los entrevistadores pueden elegir según su preferencia cuales serán los médicos a los que entrevistará, sus posibilidades de elección están restringidas a cumplir la cuota asignada.
En las encuestas políticas, el encuestador puede verse limitado a elegir personas de sexo femenino, de 30 a 45 años, que en las elecciones pasadas hayan votado por el partido X (a otros encuestadores se les encargarán otras cuotas). No hay azar, porque el encuestador elegirá a quienes prefiera o tenga más a su alcance, pero deberá respetar las cantidades que le fueron indicadas.

Muestreo de juicio o intencional
En este tipo de muestreo, conocido también como "selección experta", es el investigador quien decide qué elementos son los más adecuados para realizar la investigación. La elección se basa en la apreciación subjetiva del investigador sobre la representatividad de los elementos que muestrea. En algunos ámbitos de la investigación, suele solicitarse a expertos que seleccionen los elementos de la población que consideren más adecuados para construir la muestra, esta selección se apoya en la experiencia de los consultados.
Un primer uso de este tipo de muestreo es aquel en el que el interés no se centra en la representatividad sino en la riqueza de contenidos que pueden ofrecer algunos individuos a diferencia de otros.
Si se pretende reconstruir la historia de un asentamiento urbano marginal, sería más adecuado elegir a personas que hayan residido en él por largo tiempo. Se espera que los pobladores más antiguos puedan dar mayor información acerca de la génesis y evolución que los que han arribado recientemente. En este caso, el investigador se preocupará primero por identificar a los residentes en el asentamiento y, a partir de ese conocimiento, decidirá a quiénes va a incluir en la muestra.
Aquí no se pretende la generalización de los resultados, ni se busca encontrar una historia típica en el sentido de que, en mayor o menor grado, la mayoría de los pobladores la comparta. Lo que se busca es profundizar el análisis a partir del testimonio de aquellos a los que consideramos como informantes más idóneos.

Muestreo autoelegido

Usado principalmente en razón de la reducción de costos que implica, cuando se usa este tipo de muestreo no se selecciona a los sujetos del estudio, sino que se solicita a las personas que participen voluntariamente. La decisión de formar parte o no de la muestra queda en manos de la persona invitada, por eso se llama autoelegido, es ella quien optará por participar o no hacerlo (por falta de interés, de tiempo, etc.).

Supongamos que se pretende estudiar el gasto por familia en actividades de recreación y para reducir costos, en lugar de enviar encuestadores, se envían por correo formularios anónimos a los domicilios de los hogares seleccionados aleatoriamente. Podría suceder que aquellos hogares que tienen gastos en recreación muy bajos contesten en proporción menor que los que más gastan. De esta manera, las respuestas resultantes darían un valor promedio por encima del verdadero por el hecho que, inadvertidamente, la muestra estaría representando en mayor medida a los hogares que tienen los más altos gastos en recreación.

En el estudio acerca del comportamiento sexual de las mujeres en Estados Unidos, el Informe Hite (1976) basó sus conclusiones sobre el 3% de mujeres que remitieron completos los cuestionarios que habían recibido por correo. Aún cuando el número de cuestionarios respondidos sea alto (gracias a una numerosa muestra inicial), resulta riesgoso extender las conclusiones a la población en general. Es así porque no se sabe si ese pequeño porcentaje de mujeres que respondieron tiene conductas semejantes o muy diferentes de aquellas que no lo hicieron.

En la actualidad ya no se envían formularios por correo postal para que la gente responda, pero sí se hacen encuestas por mail o se invita a expresar la opinión en encuestas en sitios web. A esas encuestas responde una fracción muy específica de la población: la que accede a internet, que visita esa página, que tiene tiempo e interés por responder a las preguntas que allí se hacen. No es posible saber qué recorte de la población es representado por esa muestra y es completamente inválido extender los resultados a toda la población, como lo es generalizar a los visitantes de esa página, ya que no todos tienen el tiempo ni el interés para responder. Se trata de un dato imposible de adjudicar a alguna población identificable, es simplemente “la opinión de los que opinaron”. Sin embargo a veces se tratan equivocadamente esos resultados como si representaran una tendencia de opinión.

El muestreo autoelegido tiene una aplicación en el campo de la Psicología experimental cuando se trabaja con sujetos humanos. Allí, la selección de quienes participarán de los experimentos, muy raramente puede hacerse al azar, habitualmente resulta de la voluntad de personas que se prestan para los estudios. Sin embargo, es un requisito indispensable en los experimentos que se asignen las personas de manera aleatoria a los grupos experimental y control. Es decir, que se elija aleatoriamente a cuáles de las personas voluntarias se someterá a los tratamientos que se ponen a prueba y a cuáles se tomará como controles. También es frecuente usar cuotas en la asignación de sujetos a los dos grupos de modo que estén equilibrados respecto de algunas variables, cuya determinación depende del estudio particular. En este caso las cuotas indican qué proporción de casos se requieren en cada categoría de las variables de interés, la selección de los sujetos que constituirán el grupo experimental y el grupo control se hace aleatoriamente. Este tema se desarrolla con detalle en Metodología de la Investigación.

Muestreo accidental o según disponibilidad

Esta modalidad cuenta con sustancial difusión, en particular por los medios de comunicación, y consiste en entrevistar a los individuos que se encuentran accidentalmente en determinado lugar. La expresión "accidental" se presta a equívocos ya que puede confundirse con "aleatorio". Si bien no resulta posible identificar a quien toma la decisión por la selección de la muestra, nos encontramos en una situación en que múltiples factores son los responsables de la misma. Algunos de estos factores son: la voluntad de los encuestados para prestarse o no a participar de la entrevista, la elección (consciente o no) que el entrevistador hace entre las personas que transitan, el lugar donde se hace la entrevista (ciertas personas difícilmente transitan por determinadas áreas de una ciudad), etc.

El hecho que se trate de la confluencia de múltiples factores fortuitos, lo que determina cuales serán las personas que formarán parte de la muestra y quienes no, no puede considerarse sinónimo de aleatoriedad. Por lo tanto, no se trata de un muestreo probabilístico dado que no es posible asignar probabilidades a priori de pertenecer a la muestra a los elementos de la población.

En consecuencia, las conclusiones obtenidas a través de este tipo de relevamiento (porcentaje de opiniones a favor o en contra de un tema cualquiera, etc.) sólo pueden ser válidas para aquellas personas que fueron parte integrante de la muestra, no pudiendo generalizarse los resultados más allá de la misma.

Por el contrario cuando se lo utiliza en ámbitos académicos, suelen tomarse los resguardos mencionados antes para conseguir que sea tan representativo como sea posible. Para hacer esto se consideran en primer lugar las características conocidas de la población y se las reproduce en la muestra por ejemplo la proporción de varones y mujeres, los niveles de educación, como en el caso del muestreo por cuotas. Luego se busca la mayor heterogeneidad posible y un número elevado de casos. Con estos cuidados, este tipo de muestreo puede usarse, por ejemplo para la validación de escalas psicométricas.

Resumen de las técnicas de muestreo descriptas en este capítulo

Probabilísticos -Selección aleatoria -Permiten conocer el error en la generalización de los resultados	Irrestricto aleatorio o Aleatorio simple	
	Sistemático	
	Estratificado	Con afijación igual o uniforme Con afijación proporcional
	Por conglomerados o racimos	En una etapa En dos o más etapas

No probabilísticos -Selección voluntaria -Según el tipo, permiten generalizar aproximadamente los resultados -No se conoce el error de estimación	Por cuotas
	De juicio o intencional
	Autoelegido
	Accidental o por disponibilidad

Actividad práctica de repaso 7

1. En una investigación de Psicología Política se seleccionan a los encuestados en función de datos del mismo, por ejemplo sexo, edad y al preguntar la ocupación se sigue o no con la toma del cuestionario, de acuerdo a la respuesta que se dé. ¿De qué tipo de muestreo se trata?

2. Si tenemos una población homogénea, ¿cuál es el método de muestreo más?

3. ¿Cuáles son las técnicas de muestreo que permiten generalizar a la población de estudio las conclusiones obtenidas de la muestra?

4. Si decimos: "Conjunto de unidades de análisis que son objeto de un estudio particular", hacemos referencia a:

6. ¿Qué procedimiento es alternativo al muestreo irrestricto aleatorio?

7. La expresión $n_i = \frac{n}{k}$
Donde:
n_i es el tamaño de la muestras extraídas de cada estrato
n es el tamaño de la muestra total
k representa el número de estratos en que fue dividida la población

¿A qué técnica de muestro se refiere?

8. En un estudio sobre intención de voto[57] se detalla lo siguiente:
"Selección de la muestra estadística: Selección al azar de las manzanas de viviendas, para lo cual se fijó una cuota de cuatro encuestas por manzana de viviendas. Dentro de la manzana de viviendas se seleccionó al azar las viviendas donde debería de efectuarse las encuestas, y dentro de la vivienda se seleccionó a la persona encuestada respetando la cuota por sexo y edad."

¿Qué combinación de técnicas de muestreo se empleó?

9. En un estudio de intención de votos de la Provincia de Buenos Aires, realizado por la Universidad Abierta Interamericana se aclara:

[57] Disponible en: http://www.cpi.com.pe/descargas/OPLI20061009.pdf

Población Target: Residentes en la Provincia de Buenos Aires, en condiciones de votar.
Tamaño Muestral: 1273 casos efectivos.
Tipo de Muestra: Aleatoria Simple, ponderada por cuotas de Género, Rango de Edad y Nivel Socio Económico (18 segmentos).
Margen de Error Muestral: +/-2.7% (Para P = 0.50).
Instrumento de Recolección de Datos: Cuestionario Cerrado.
Fecha del Campo: 22 al 25 de Junio, 2009.

Se han utilizado como técnicas de muestreo:

10. La Estadística Inferencial nos permite efectuar generalizaciones utilizando los datos que nos proporciona la Estadística Descriptiva. Para ello, los muestreos deben cumplir ciertas características. Existen dos grandes divisiones entre las técnicas de muestreo, caracterice a cada una de ellas.

11. Determine en cada uno de los siguientes casos la unidad de análisis y la técnica de muestreo que se usa.
a. diez escuelas y, dentro de cada una, se seleccionan, también al azar, cinco docentes
b. diez escuelas y en cada una se entrevista a todos los docentes
c. diez escuelas públicas, diez privadas laicas, diez privadas religiosas, en cada una se mide la proporción de alumnos que repiten
d. diez escuelas y se releva la cantidad de alumnos por docente

Capítulo 8: Distribuciones en el muestreo

Eduardo Bologna

En el capítulo sobre muestreo indicamos la importancia del carácter aleatorio en la elección de las unidades que constituirían la muestra. Esta exigencia se funda en la necesidad de generar variables aleatorias cuya distribución de probabilidad conozcamos, o bien podamos suponer. Si podemos hacer eso, conoceremos las probabilidades asociadas a determinados resultados muestrales y podremos entonces usarlos para inferir sobre la población, que es, no lo olvidemos, nuestro objetivo. Este capítulo es breve y va a ligar lo que traemos hasta aquí sobre probabilidad y muestreo, a fin de mostrar cómo usar esos conceptos para hacer estimaciones.

En todo lo que queda de la materia trataremos con datos que provienen de muestras probabilísticas, es decir, unidades de análisis que fueron seleccionadas usando alguno de los procedimientos de muestreo probabilístico que se trataron en el capítulo 7. En consecuencia, las medidas que calculemos a partir de esos datos dependerán del azar. Las expresiones que presentaremos para hacer inferencias corresponden al caso de muestreo irrestricto aleatorio, cuando se usen otros diseños de muestra (siempre probablísticos), será necesario introducir correcciones; el texto de Scheaffer et al (1987) es una adecuada referencia para esas correcciones.

Por ejemplo; si el conjunto completo de estudiantes de Psicología (la población estudiantil) tiene una edad promedio de 25,4 años, cuando extraiga una muestra de 400 alumnos al azar, la edad promedio puede ser un número muy diferente de 25,4 años. Recordemos que las muestras tienen carácter aleatorio, no podemos asegurar que la variable que estamos estudiando se asemeje en la muestra al valor que tiene en la población, a menos que conozcamos éste último, en cuyo caso no estaríamos preocupados por hacer inferencias desde la muestra. Por eso, el resultado que obtengamos en la muestra es un resultado que depende del azar: resulta ser una variable aleatoria. La media muestral, por ser una medida calculada sobre los valores obtenidos en una muestra aleatoria, es una variable aleatoria, su valor depende de cuáles sean los casos que constituyen la

muestra y esto depende del azar, que es lo que hemos pedido como requisito al procedimiento de muestreo.
Estamos entonces refiriéndonos a entidades diferentes cuando hablamos de la media poblacional y la media muestral, aunque el procedimiento de cálculo sea el mismo. La media poblacional es un valor que puede ser conocido (si hemos observado a toda la población) o desconocido, pero en todo caso es fijo. Cuando hacemos estimaciones no sabemos cuánto vale la media de la población, pero sí sabemos que es un número estable, fijo. Por el contrario, la media muestral —y porque depende de los casos que hayan sido aleatoriamente seleccionados para constituir la muestra— depende finalmente del azar, luego es una variable aleatoria.

Las medidas que se refieren a la población se denominan **parámetros** (a veces se dice parámetros poblacionales, para acentuar que se trata de la población), por ejemplo, el valor de 25,4 años que mencionamos más arriba es la media paramétrica o también media poblacional. Si se trabaja con variables cualitativas, se tratará de la proporción de alguna categoría (proporción de varones, proporción de afectados de cierta patología, de egresados, de repitentes, etc.), se tratará allí de la proporción paramétrica o proporción poblacional. O bien de la varianza, la desviación estándar o también de coeficientes de correlación (Pearson, Spearman). Todas las medidas que hemos mencionado hasta aquí pueden calcularse sobre la población completa si se hace un censo, o bien sobre una muestra que debe ser aleatoria si se quiere luego hacer inferencias. En todos los casos que se trate de medidas que se refieran a la población completa, las llamaremos paramétricas. Solo podrán ser conocidas en los casos en que la población íntegra sea observada, cuando se haga un relevamiento exhaustivo, un censo. En general esto no es posible: hay poblaciones que son inaccesibles o de muy alto costo para alcanzarlas totalmente. Cuando el objetivo de un estudio es, por ejemplo, conocer si una droga tiene efectos para aliviar síntomas de la depresión, deseamos que el resultado que hallemos pueda ser válido para todas las personas diagnosticadas de depresión. Este conjunto incluye también a quienes serán diagnosticados de depresión en el futuro; a quienes, evidentemente, no podemos observar ahora: ya nos hemos referido a esto como poblaciones hipotéticas. Por eso la única alternativa es extraer una muestra, observar sobre ella lo que nos interesa (el efecto de la droga) y luego generalizar el resultado. Otro ejemplo: si nuestro objetivo es ver el efecto de la educación sobre la cantidad de hijos que tienen las mujeres

de la ciudad de Córdoba, es difícil que tengamos presupuesto para interrogar a todas las mujeres de la ciudad. Tenemos como alternativa usar datos de un censo, pero éstos se hacen cada diez años y puede que el más cercano ya esté desactualizado. Nuevamente allí, optaremos por extraer una muestra aleatoria de mujeres que viven en la ciudad y generalizaremos nuestros hallazgos a partir de ella. Lo que obtengamos en la muestra depende del azar, porque fue el azar el que determinó qué mujeres participaron de la muestra.

Por eso, lo que nos interesa en este capítulo es concentrarnos en las situaciones en que no conocemos los parámetros de una población, porque entonces deberemos **estimarlos** desde la muestra. Diremos que buscamos estimar diferentes parámetros. Ese es el tema que tratará el próximo capítulo: "Estimación de Parámetros". También puede suceder que haya algún valor hipotético para un parámetro, al que necesitemos poner a prueba. En ese caso usaremos los datos de la muestra para **probar una hipótesis**, de eso trata un amplio campo que trataremos más adelante llamado "Pruebas de Hipótesis".

Las medidas calculadas sobre los datos de la muestra se denominan **estadísticos** (a veces aparecen como estadísticos muestrales, nuevamente para acentuar de dónde provienen). Hablaremos entonces de la media muestral, la proporción, la varianza o el coeficiente de correlación muestral. Los valores que de ellos obtengamos en la muestra nos permitirán estimar los correspondientes valores paramétricos. Por eso decimos que la media muestral es el estimador de la media poblacional y del mismo modo con la proporción, la varianza, la desviación estándar o los coeficientes de correlación.

Para distinguir entre las medidas de la población y las de la muestra usaremos diferentes símbolos. De manera general, las letras latinas se usarán para identificar medidas descriptivas obtenidas sobre datos muestrales, mientras que usaremos letras griegas para referirnos a los valores de la población. Pero esto no es siempre así, por razones de tradición en el uso, en el caso de la proporción se distingue la poblacional de la muestral usando P (mayúscula) para la primera y p (minúscula) para la segunda[58]. La siguiente es la notación que usamos y la correspondencia entre valores poblacionales y muestrales.

[58] Dado que puede haber confusión al escribir p ó P, es frecuente usar $\hat{p}$ para indicar la proporción muestral.

	En la población	En la muestra
Cantidad de casos	N	n
Media	μ (mu)	$\bar{x}$
Proporción	P	$\hat{p}$
Varianza	σ^2 (sigma cuadrado)	s^2
Desviación Standard	σ (sigma)	s
Coeficiente de correlación de Pearson	ρ (rho)	r
Coeficiente de correlación de Spearman	ρ_s (rho ese)	r_s
Pendiente de la recta de regresión	β_1 (beta uno)	b_1
Ordenada al origen de la recta de regresión	β_0 (beta cero)	b_o

Las operaciones que se realizan para calcular los estadísticos que hemos tratado hasta aquí son las mismas que para los parámetros, solo difieren los nombres de cada elemento. Así para calcular μ debemos hacer como con $\bar{x}$: sumar todos los valores y dividir por el total de casos. En lugar de:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * f_i}{n}$$

escribiremos:

$$\mu = \frac{\sum_{i=1}^{N} x_i * f_i}{N}$$

que solo difiere en el nombre de la media y la cantidad de casos.

Una excepción a esto es el caso de la varianza, porque el modo de calcularla es levemente distinto si se trabaja con datos muestrales o poblacionales. Cuando es calculada sobre los elementos de la muestra, la varianza muestral es, según vimos en el capítulo 3:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Pero cuando la varianza se calcula sobre los elementos de la población, su cálculo se realiza con la expresión:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

En esta última expresión cambiaron los nombres de los elementos con los que se opera (μ en lugar de $\bar{x}$ por, N en lugar

de *n*), y además se ha eliminado el -1 del denominador. Ya no es el total de casos menos uno, sino el total de casos. El origen de esto es que esta última expresión (con denominador *N*) es la definición original de la varianza, mientras que la que usamos en la muestra (s^2 con denominador *n-1*) es una corrección, que se hace para que sea un buen estimador de la varianza poblacional.[59]
Con InfoStat® el cálculo de la varianza puede solicitarse para la población o para una muestra, la notación que usa ese programa es *Var(n)* para referirse a la varianza poblacional, es decir a σ^2, y *Var(n-1)* para indicar la varianza muestral, s^2, en referencia al denominador que se usa para calcular cada una.

Una vez acordados los elementos con los que trabajaremos en este capítulo, debemos volver sobre el problema del carácter aleatorio de los estimadores (los estadísticos muestrales). Como ya tratamos en el capítulo 6 las variables aleatorias tienen distribuciones de probabilidad asociadas y eso es lo que permite calcular la probabilidad de ocurrencia de sus diferentes valores. A los fines de este capítulo, conocer las distribuciones de probabilidad de los estadísticos muestrales nos permitirá ponerlos en relación con los parámetros a los que estiman: usaremos las distribuciones de probabilidad que hemos mencionado en el capítulo 6 para ligar los estimadores (conocidos) con los parámetros (desconocidos).

Evitaremos los desarrollos matemáticos que llevan a establecer dicha relación, en cambio, empezaremos viendo esto con un ejemplo que trata sobre una muy pequeña población ficticia. Supongamos que sea la de tres pacientes de un hospital psiquiátrico que los tiene como únicos internados[60]. Llamaremos A, B y C a estos tres pacientes y consideraremos dos variables: el *tiempo que llevan desde su ingreso* (expresado en meses) y el *sexo*. La primera variable es cuantitativa, la

[59] Lo que se consigue con esa corrección es que, si se calculan las varianzas de todas las muestras que se saquen de una población dada, éstas promedien la varianza de la población, lo cual no sucede si se usa el denominador *n*. Como lo mencionamos en el capítulo 7, ésta es la que llamamos insesgabilidad del estimador, por lo que podemos decir que el denominador de la varianza muestral se transforma en *n-1* para lograr que sea un estimador insesgado de la varianza poblacional.
[60] Suponer que en el hospital solo hay tres internados es una gran simplificación, por cierto ficticia. La usamos solo con fines expositivos, para poder hacer la comparación entre la población y las muestras.

segunda cualitativa. La siguiente es la matriz de datos para esta población ficticia y las variables indicadas.

caso	Paciente	Meses desde el ingreso	Sexo
1	A	3	Varón
2	B	4	Mujer
3	C	5	Varón

Para describir esta población indicaremos la media y la varianza de la variable cuantitativa (meses desde el ingreso):

$$\mu = \frac{3+4+5}{3} = 4$$

$$\sigma^2 = \frac{(3-4)^2 + (4-4)^2 + (5-4)^2}{3} = \frac{2}{3} = 0{,}67$$

Además, calcularemos —también sobre datos de la población completa— la proporción de mujeres[61], codificando como 1 la presencia de un "éxito" y 0 su ausencia:

$$P = \frac{0+1+0}{3} = \frac{1}{3} = 0{,}33$$

Estos valores (μ, σ^2 *y* P) son los que llamamos parámetros y caracterizan a las dos variables para la población completa.
Ahora veremos qué sucede cuando se muestrea. Para ello vamos a sacar todas las muestras posibles de tamaño dos (podrían haber sido de otro tamaño, lo elegimos para este ejemplo), y las muestras que extraigamos serán con reposición[62]. Así resultan las siguientes nueve muestras: AA, AB, AC, BA, BC, BB, CA, CB, CC. Cada una de ellas tiene probabilidad 1/9 de ser seleccionada aleatoriamente.
En cada una de las muestras calcularemos la media del tiempo de internación ($\bar{x}$) y la proporción de mujeres ($\hat{p}$).

[61] Esta categoría se eligió arbitrariamente, podría haber sido la otra y calcular la proporción de varones.

[62] Esta forma de extraer una muestra no es la que se aplica en la práctica, ya que no es correcto seleccionar dos veces al mismo individuo, como lo mencionamos en el capítulo 8, ya que atenta contra la heterogeneidad de la muestra. Sin embargo, este muestreo con reposición en una buena aproximación a los muestreos verdaderos (sin reposición) cuando se trabaja con poblaciones grandes. Es así porque la extracción de unos pocos casos, aunque sean sin reposición, altera poco la probabilidad de ser extraídos de los otros casos.

Muestra	Meses de internación	$\bar{x}$	Cantidad de mujeres	$\hat{p}$
AA	3-3	3,0	0	0,0
AB	3-4	3,5	1	0,5
AC	3-5	4,0	0	0,0
BA	4-3	3,5	1	0,5
BB	4-4	4,0	2	1,0
BC	4-5	4,5	1	0,5
CA	5-3	4,0	0	0,0
CB	5-4	4,5	1	0,5
CC	5-5	5,0	0	0,0

Antes de analizar estos resultados, los organizaremos mejor, presentando las distribuciones de frecuencia de las $\bar{x}$ y de las $\hat{p}$:

$\bar{x}$	f	f
3,0	1	0,11
3,5	2	0,22
4,0	3	0,33
4,5	2	0,22
5,0	1	0,11
Total	9	1,00

$\hat{p}$	f	f
0,0	4	0,44
0,5	4	0,44
1,0	1	0,11
Total	9	1,00

Distribución de la media muestral

Vamos a concentrarnos inicialmente en la distribución de $\bar{x}$. Lo primero que se observa en la tabla de distribución de frecuencias de $\bar{x}$ es lo que significa que $\bar{x}$ sea una variable: quiere decir que las diferentes muestras ofrecen valores diferentes de $\bar{x}$; o bien, que el valor de $\bar{x}$ varía dependiendo de cuál haya sido la muestra que se seleccionó al azar. En segundo lugar, estos valores no se refieren a casos individuales sino a promedios obtenidos en muestras de tamaño dos.

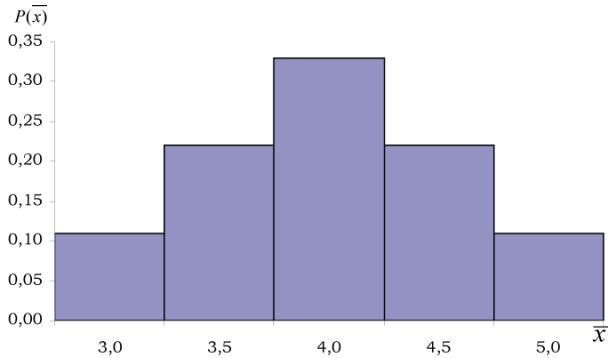
Si se usan las frecuencias relativas como aproximaciones a la probabilidad, leemos en la tabla que la probabilidad de obtener una muestra en la que el promedio sea de tres meses de internación es de 0,11. Expresamos estos simbólicamente así:

$$P(\bar{x} = 3) = 0{,}11$$

Comparando las probabilidades de los diferentes valores de $\bar{x}$, se puede ver que resulta más probable hallar una media de 4 que de 3,5 ó de 3; porque $P(\bar{x} = 4) = 0{,}33$ mientras que $P(\bar{x} = 3) = 0{,}11$. Se trata de un resultado importante, porque la media de la población es 4, por lo que vemos que es más probable encontrar una media muestral cerca de la media poblacional que lejos de ella.

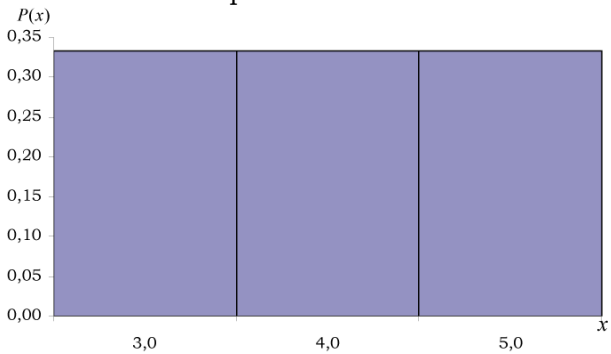
La representación gráfica de la distribución también nos aporta información sobre la relación entre $\bar{x}$ y μ.

Gráfico 1: Distribución de probabilidades de $\bar{x}$



Vemos aquí que $\bar{x}$ asume valores cuya distribución es simétrica alrededor del valor 4, que es la media poblacional (μ). Resulta interesante comparar esta distribución con la que tiene la variable original (el número de meses de internación de cada paciente):

Gráfico 2: Distribución de probabilidades de x



Observemos primero el nombre de los ejes: el Gráfico 1 muestra la distribución de las $\bar{x}$: el promedio de los tiempos de internación de muestras de dos pacientes. Son nueve casos, cada uno es una muestra de tamaño 2. Sus valores están en el eje horizontal y sus probabilidades se indican como $P(\bar{x})$ en el eje vertical.

El Gráfico 2, representa los valores de x: los tiempos de internación de cada paciente individualmente. Son tres casos, cada uno es un paciente. Esos son los valores que se ubican en el eje horizontal, las probabilidades del eje vertical son $P(x)$.

La del Gráfico 2 es la distribución en la población, en este caso es uniforme[63], pero podría tener cualquier otra forma. Por el contrario, la del Gráfico 1 es la distribución de las medias muestrales; es simétrica y va a tender a tener esta forma a medida que las muestras sean de mayor tamaño, independientemente de cuál sea la distribución que la variable tenga en la población. Luego volveremos sobre esta importante observación.

Vamos ahora a describir esta nueva variable $\bar{x}$, a través de su esperanza y su varianza. Conviene regresar al capítulo 6 para recordar el significado de estas medidas.
Calculemos primero la esperanza de $\bar{x}$, aunque, por la forma de la distribución, ya deberíamos conocerla:

$$E(\bar{x}) = 3 * 0{,}11 + 3{,}5 * 0{,}22 + 4 * 0{,}33 + 4{,}5 * 0{,}22 + 5 * 0{,}11 = 4$$

Encontramos que el valor esperado para las medias muestrales coincide con la media de la población. Este es el primer resultado que nos interesa retener de la relación entre $\bar{x}$ y μ; la expresamos como:

$$E(\bar{x}) = \mu$$

Esto implica que, si de una población se extraen todas las muestras posibles de un determinado tamaño, y en cada una de ellas se calcula la media, el promedio de esas medias muestrales coincide con la media de la población completa. Esta cualidad según la cual la esperanza del estimador es igual al parámetro, como ya hemos dicho, se llama insesgabilidad. Cuando un estimador cumple con esa condición (como sucede con $\bar{x}$ como estimador de μ) se dice de él que es insesgado.

La media muestral ($\bar{x}$) es un estimador **insesgado** de la media poblacional (μ) porque su esperanza es igual al parámetro que estima.

Ahora vamos a calcular la varianza de $\bar{x}$:

$$V(\bar{x}) = (3-4)^2 * 0{,}11 + (3{,}5-4)^2 * 0{,}22 + (4-4)^2 * 0{,}33 + (4{,}5-4)^2 * 0{,}22 + (5-4)^2 * 0{,}11 = 0{,}33$$

Este valor no es el mismo que el de σ^2, sino que es la mitad. Se trata nuevamente de un resultado que podemos generalizar y depende del tamaño de la muestra. En el caso de nuestro ejemplo, las muestras son de tamaño dos y por esa razón la

[63] La primera distribución que mencionamos en el capítulo 6.

varianza de la media muestral es la varianza poblacional dividida por dos. Si hubiésemos tomado muestras de tamaño 3, la varianza habría quedado reducida a la tercera parte. Para otro tamaño de muestra, la varianza de $\bar{x}$ será diferente. De modo general, la varianza de $\bar{x}$ se relaciona con la varianza de x (σ^2) a través de:

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

De manera alternativa, indicamos a esta varianza como $\sigma_{\bar{x}}^2$, con lo que:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Esta fórmula indica que la varianza de las $\bar{x}$ varía inversamente con el tamaño de la muestra.
Este también es un resultado muy importante: nos dice que cuanto más grande sea la muestra tanto más pequeña será la varianza de las medias muestrales. Recordemos que la varianza mide la dispersión de los valores de una variable, por lo que esta varianza mide la dispersión entre los valores de las diferentes medias muestrales. El hecho que disminuya cuando aumenta el tamaño de la muestra significa que para muestras más grandes, las medias muestrales tendrán menos dispersión, es decir que serán más similares entre sí. En el capítulo 7 llamamos a esta propiedad consistencia, entonces este resultado puede expresarse indicando que la media muestral es un estimador consistente de la media poblacional, porque su varianza se reduce a medida que aumenta el tamaño de la muestra.

La media muestral ($\bar{x}$) es un estimador **consistente** de la media poblacional (μ) porque su varianza disminuye cuando se toman muestras de mayor tamaño.

De la varianza surge inmediatamente la desviación estándar de $\bar{x}$ es $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Es común referirse a ella como el **error estándar de la media**, y suele abreviarse, en los programas de análisis de datos como EE.
Este es el segundo resultado de importancia sobre la relación entre $\bar{x}$ y μ.

En la realidad, fuera de este ejemplo simplificado, solo extraemos una muestra y en ella calculamos la media, lo que hemos visto hasta aquí nos dice que nuestra media muestral tiene más chances de estar cerca de la media poblacional cuanto

más grande sea la muestra que tomemos. Esto es así porque a medida que aumenta el número de casos en la muestra (n) disminuye la varianza y en consecuencia las medias muestrales tienen más probabilidad de estar cerca de la media poblacional. Por eso, será más probable que la única media muestral con que contamos sea próxima al parámetro que estamos estimando.
Destaquemos que, si bien no es seguro que la media muestral sea cercana a la poblacional, sí sabemos que es menos probable que esté lejos cuando la muestra es más grande.

La tercera característica de la distribución de la media muestral se refiere a su forma. En el Gráfico 1 vimos que las $\bar{x}$ del ejemplo alcanzaron una distribución unimodal y simétrica, sin importar que en la población la variable hubiese tenido una distribución uniforme. Este resultado también es general y es más amplio aun: a medida que aumenta el tamaño de las muestras, la distribución de las medias muestrales tiende a ser normal. La prueba de esta afirmación constituye el Teorema Central del Límite, cuyo enunciado puede resumirse como: una suma de n variables aleatorias tiende a tener distribución normal a medida que aumenta n, independientemente del modo en que esté distribuida esa variable. Para nuestro uso, la media es una suma de valores de una variable aleatoria (porque esos valores provienen de una muestra aleatoria) dividida en el total de casos. Por eso el teorema nos dice que si se trata de muestras grandes, puede tratarse a $\bar{x}$ como teniendo una distribución normal.
Volvemos sobre esta idea porque es de gran importancia para todo lo que viene a continuación. En el capítulo 6 presentamos el experimento de tirar un dado dos veces y contar la suma de los puntos obtenidos. Definimos la variable aleatoria S que resulta de la suma de dos variables aleatorias, porque cada tirada del dado da lugar a un resultado entre 1 y 6 que depende del azar. La variable que resulta de tirar una sola vez el dado tiene distribución uniforme: todos los números tienen la misma probabilidad 1/6. La suma de puntos de los dos dados, a la que llamamos S, ya no tiene distribución uniforme, porque es más probable obtener 7 (que puede resultar de 6+1, 5+2, 4+3, 3+4, 2+5 ó 1+6) que un 12 (que solo puede obtenerse con 6+6).

Gráfico 3: Distribución de probabilidades de la variable x = *puntaje obtenido al tirar un dado*

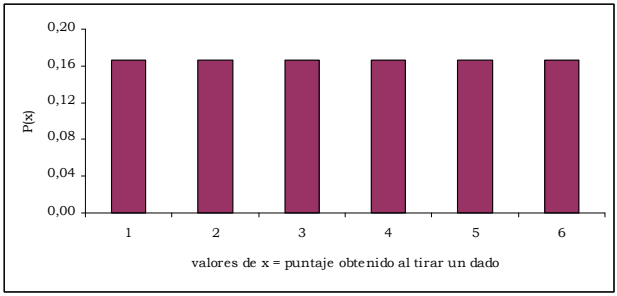
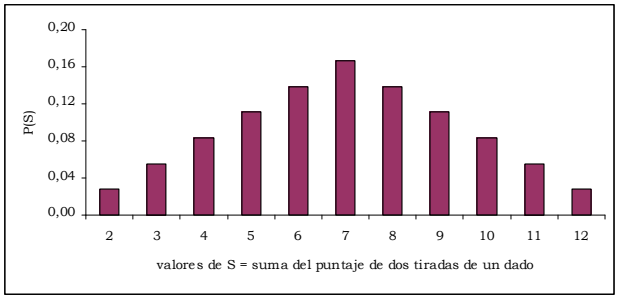


Gráfico 4: Distribución de probabilidades de S = *suma de puntajes obtenidos al tirar dos veces un dado*



Si bien la distribución de esta variable no es normal, podemos ver la tendencia, "va camino a ser normal". En este caso la muestra tiene solo dos casos (las dos tiradas del dado); cuando son más casos, la distribución va volviéndose más cercana a la normal.

Esto es lo que sucede cuando se extraen muestras de una población: la distribución de la variable en la población puede tener cualquier distribución (en el caso del dado es uniforme), pero a las medias muestrales les sucede lo mismo que a S, van tendiendo a tener una distribución normal a medida que se toman muestras de mayor tamaño. Éste es el tercer resultado que necesitamos para relacionar a la media muestral con la poblacional y poder hacer las primeras inferencias.

Resumimos en el recuadro siguiente los tres resultados mostrados hasta este punto.

$$E(\bar{x}) = \mu$$

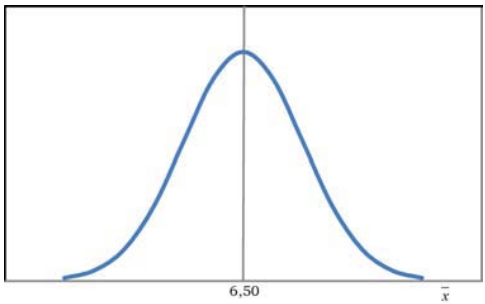
$$V(\bar{x}) = \frac{\sigma^2}{n} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \to N\left(\mu, \frac{\sigma^2}{n}\right) \; cuando \; n \to \infty$$

La última expresión sintetiza las anteriores y agrega que la distribución de $\bar{x}$ es normal.
Entonces, del recuadro leemos que:

- La esperanza de $\bar{x}$ es la media de la población.
- La varianza de $\bar{x}$ es la varianza de la población dividida en el tamaño de la muestra[64].
- La media muestral tiende a tener una distribución normal con media μ y varianza $\frac{\sigma^2}{n}$, cuando el tamaño de la muestra aumenta.

Veamos una aplicación de este resultado. Supongamos que en la población de adultos la media de horas diarias de sueño fuera de 6,5 hs. ($\mu = 6{,}5$) con varianza de 9 hs^2. ($\sigma^2 = 9$). Si eligiéramos al azar muestras de 200 personas (n=*200*) y registráramos el número promedio de horas de sueño en cada muestra, encontraríamos que la distribución de esas horas promedio sería como la siguiente:



[64] De modo equivalente, la desviación estándar de la media es la desviación estándar de la población dividida en la raíz cuadrada del tamaño de la muestra. Es conocida también como error estándar de la media.

En el eje horizontal están graficados los valores de las posibles $\bar{x}$ que resultarían si sacáramos muestras aleatorias de esa población. La distribución normal está centrada en 6,50, que es la media de la población. El gráfico muestra con claridad que "lo más probable" es encontrar a $\bar{x}$ en los alrededores de 6,50.
Con la media y la desviación estándar de esta variable $\bar{x}$ podemos calcular un puntaje *z*, lo que conocíamos como desvío estándar. Para llegar a la forma que tiene ahora *z*, recordemos que lo definimos como el número de desviaciones estándar a las que la variable se encuentra de la media, por eso era la diferencia entre el valor de la variable y el de la media, dividida en la desviación estándar:

$$z = \frac{x - \bar{x}}{s}$$

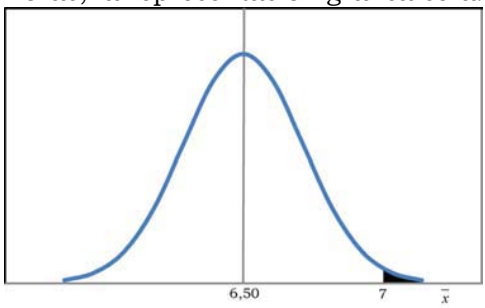
La diferencia ahora es que nuestra variable es $\bar{x}$, su media es μ y su desviación estándar es el que hemos llamado error estándar de la media: $\sigma_{\bar{x}}$. Con lo que el puntaje *z* asociado a cada valor de $\bar{x}$, tiene ahora la forma:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Si se reemplaza el error estándar de la media por su valor, se obtiene:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Si ahora nos preguntamos por la probabilidad de encontrar muestras de 200 personas cuyos promedios de horas de sueño superen las 7 horas, la representación gráfica será:



Y el valor de esa probabilidad se calcula observando a cuántas desviaciones estándar es encuentra de la media, es decir, calculando el puntaje z correspondiente a la media de 7 horas:

$$z = \frac{7 - 6{,}5}{\frac{3}{\sqrt{200}}} = \frac{0{,}5}{0{,}21} = 2{,}38$$

Pedimos a InfoStat® la probabilidad de superar este valor[65] *P(z>2,38)* y obtenemos 0,009. La lectura de este resultado es que solo el 0,9% de las muestras de tamaño 200 darán media muestral mayor a 7hs.
Para ver el efecto del tamaño de la muestra repitamos el cálculo con muestras de 30 casos[66]. El puntaje z es ahora:

$$z = \frac{7 - 6{,}5}{\frac{3}{\sqrt{30}}} = \frac{0{,}5}{0{,}55} = 0{,}91$$

Y *P(z>0,91)= 0,181* es decir que el 18% de las muestras de tamaño 30 darán media que supere a 7hs. Esto indica que cuando las muestras son de mayor tamaño es menor la probabilidad de encontrar valores en los extremos de la distribución.

Supongamos que la variable *horas de sueño* tiene distribución normal en la población y calculemos la probabilidad de encontrar personas que duerman más de 7 horas. Antes destaquemos la diferencia entre esta pregunta y las dos anteriores: las probabilidades que calculamos antes son las de hallar muestras de 200 personas o de 30 personas con promedio superior a 7 horas de sueño, ahora preguntamos por la probabilidad de encontrar individuos que tengan más horas de sueño que esa cifra. El puntaje z es ahora:

$$z = \frac{7 - 6{,}5}{3} = \frac{0{,}5}{3} = 0{,}17$$

La fórmula cambia porque la pregunta no es por medias muestrales sino por valores individuales, por eso no hay n[67]. La probabilidad es ahora:

$$P(z > 0{,}17) = 0{,}432$$

Que leemos diciendo que el 43% de los individuos duerme 7 horas o más. Esto indica que es mucho más probable encontrar individuos que se alejen de la media, que grupos de a 30 ó de a

[65] Alternativamente podemos ingresar al calculador de InfoStat® con la media 6,5 y la desviación estándar 0,14 y obtener el mismo resultado sin pasar por *z*. Hicimos este recorrido porque necesitamos estar familiarizados con el puntaje z para los contenidos que veremos más adelante.
[66] Es el mínimo tamaño que podemos usar para que sea válida la aplicación de la distribución normal
[67] Aunque puede interpretarse como si los individuos constituyeran muestras de tamaño 1.

200 individuos cuyo promedio se aleje de la media. Cuando las muestras son de mayor tamaño, tanto más improbable resulta encontrarlas lejos de la media, eso es lo que está expresado cuando n aparece en el denominador de la desviación estándar: mayor n implica menor dispersión, y de ello se sigue que son menos probables los casos extremos.

Distribución de la proporción muestral

El razonamiento para llegar a la relación que hay entre el estimador $\hat{p}$ y el parámetro correspondiente, P es completamente análogo al de la media, por lo que no recorreremos nuevamente los mismos pasos que llevaron a establecer la relación entre $\bar{x}$ y μ.

En primer lugar, y como sucede en el caso de la media, $\hat{p}$ es un estimador insesgado de P.

La proporción muestral $\hat{p}$ es un estimador **insesgado** de la proporción poblacional P porque su esperanza es igual al parámetro que estima.

Si extrajéramos todas las muestras posibles de una población y calculáramos en cada una la proporción de casos en una categoría de una variable, el promedio de todas esas proporciones muestrales, daría como resultado la proporción de casos que hay en esa categoría en la población. Por lo que podemos escribir que $E(\hat{p}) = P$

Acerca de la dispersión que alcanzan las diferentes $\hat{p}$ en las muestras, hay más diferencia con la media. En efecto, cuando se trata con variables cualitativas (nominales u ordinales) no hay distancias y en consecuencia no se puede usar una desviación estándar. Por el contrario, en el capítulo 3 dijimos que la dispersión de una variable de este nivel se aprecia a través de la idea de incertidumbre: habrá tanto menos dispersión cuando mayor sea la concentración de casos en una categoría de la variable. Así, la distribución:

candidato que votará	f
A	0,09
B	0,74
C	0,11
D	0,06
Total	1,00

Tiene menos dispersión que:

candidato que votará	F
A	0,23
B	0,31
C	0,26
D	0,20
Total	1,00

Porque si tuviéramos que "adivinar" quién va a ganar las elecciones, en el primer caso estaríamos más seguros de inclinarnos por el candidato B, que en el segundo. Aunque en ambas distribuciones el modo es el candidato B, en la primera la concentración es mayor y, por lo tanto menor es la incertidumbre, tenemos mayor certeza, menos dispersión.
Así es como se trata el problema de la variabilidad en variables que no admiten la medición de distancias. Pero en el caso que nos interesa, estamos calculando $\hat{p}$ como la proporción de una categoría, sin considerar cómo se distribuyen las otras, es decir que trabajamos con variables dicotómicas. Si nos concentramos en el candidato A, las dos tablas anteriores se reducen a:

candidato que votará	f
A	0,09
otro candidato	0,91
Total	1,00

candidato que votará	f
A	0,23
otro candidato	0,77
Total	1,00

¿Cuál de las dos distribuciones tiene mayor dispersión? La primera tiene una mayor concentración en la categoría "otro candidato" que la segunda, por lo que diremos que tiene menor dispersión. Por eso, en variables dicotómicas, la mayor diferencia entre las dos proporciones indica la mayor concentración y por ello, la menor dispersión.
La forma operativa de evaluar esto es multiplicando las dos frecuencias relativas: 0,09X0,91 en el primer caso, y 0,23X0,77 en el segundo. Esos productos dan 0,08 y 0,18 respectivamente; interpretamos estos valores como indicadores de la menor dispersión de la primera distribución.
La operación que hemos hecho fue la de multiplicar a la proporción de casos de una categoría por la de la otra, que se escribe como PX(1-P), o simplemente P(1-P). Esta es la medida de la dispersión en variables nominales, que reemplaza a la

varianza de las variables cuantitativas. Por analogía con lo que sucedió con la media, la varianza de la proporción muestral es la varianza dividida el tamaño de la muestra. Será entonces:

$$V(\hat{p}) = \frac{P(1-P)}{n}$$

Como antes, muestras de mayor tamaño dan lugar a menor variabilidad. También puede expresarse la varianza de $\hat{p}$ como $\sigma_{\hat{p}}^2$, con lo que:

$$\sigma_{\hat{p}}^2 = \frac{P(1-P)}{n}$$

La desviación estándar de $\hat{p}$ será:

$$\sigma_{\hat{p}}^2 = \sqrt{\frac{P(1-P)}{n}}$$

Esta expresión también se conoce como **error estándar de la proporción**.

Por último, y como también sucedió con la media, $\hat{p}$ tiende a tener distribución normal a medida que crece el tamaño de la muestra.

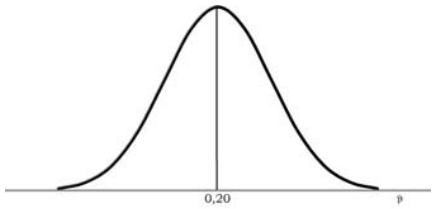
Resumimos esto en el siguiente recuadro:

$$E(\hat{p}) = P$$

$$V(\hat{p}) = \frac{P(1-P)}{n} \qquad \sigma_{\hat{p}}^2 = \sqrt{\frac{P(1-P)}{n}}$$

$$\hat{p} \to N\left(P, \frac{P(1-P)}{n}\right) \; cuando \; n \to \infty$$

Veamos una aplicación: Sea que trabajamos con la proporción de personas que consultaron a un centro de salud durante el año pasado. Supongamos que, para una población determinada, esa proporción sea del 20% (P=0,20). Si extraemos muestras de tamaño 100 y en cada una de ellas observamos la proporción de personas que consultaron a ese centro de salud durante el año pasado, la distribución de esa proporción muestral será:



El eje horizontal representa todas las proporciones que pueden encontrarse en las diferentes muestras. La distribución está centrada en el parámetro 0,20.

Si nos interesamos por la probabilidad de encontrar muestras de 100 personas en las cuales, por ejemplo, más del 25% haya consultado ese año al centro de salud, lo escribiremos como $P(\hat{p} > 0{,}25)$. Para calcular esta probabilidad debemos usar la distribución normal y para eso necesitamos el puntaje z correspondiente a ese valor de $\hat{p}$. Para calcularlo, haremos:

$$z = \frac{\hat{p} - P}{\sigma_{\hat{p}}}$$

Reemplazando por el error estándar de $\hat{p}$, tenemos:

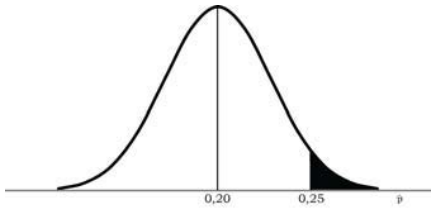
$$z = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

Esta es la expresión con la que transformamos los valores de $\hat{p}$ a puntajes z. En el ejemplo es$\hat{p} > 0{,}25$ y será:

$$z = \frac{0{,}25 - 0{,}20}{\sqrt{\frac{0{,}20(1-0{,}20)}{100}}} = \frac{0{,}05}{0{,}04} = 1{,}25$$

En InfoStat® hallamos que:

$$P(\hat{p} > 0{,}25) = P(z > 1{,}25) = 0{,}1056$$



A partir de los resultados encontrados en este capítulo podremos hacer estimaciones de la media y la proporción poblacionales a partir de los respectivos valores muestrales. Hemos visto entonces que el carácter aleatorio de las muestras hace que las estimaciones sean inciertas, pero que, debido a que conocemos la distribución de probabilidades de los estimadores podemos establecer qué valores de ellos son más probables, así podremos hacer el camino inverso, que es el que más nos

interesa: el de alcanzar a los parámetros a partir de los estimadores.

Para los dos estimadores que hemos mencionado en este capítulo se cumple la relación:

$$z = \frac{Es - Pa}{EEE}$$

En que resumimos:

Pa: el parámetro
Es: su estimador
EEE: el error estándar del estimador

Veremos en los próximos capítulos que esta expresión es válida para relacionar otros parámetros con sus correspondientes estimadores, mientras pueda usarse la distribución normal.

Con los contenidos vistos hasta este punto, en la expresión anterior, los componentes pueden ser:

Pa: μ ó P

Es: $\bar{x}$ ó $\hat{p}$

EEE: $\frac{\sigma}{\sqrt{n}}$ ó $\sqrt{\frac{P(1-P)}{n}}$

Resumen de la relación entre los estimadores y los parámetros tratados en este capítulo

Nivel de medición	Parámetro	Estimador	Varianza del estimador	Error estándar del estimador
Intervalar o proporcional	Media: μ	$\bar{x}$	$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Nominal	Proporción de casos en una categoría: P	$\hat{p}$	$\sigma_{\hat{p}}^2 = \frac{P(1-P)}{n}$	$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$

Actividad práctica de repaso 8

1. ¿Cómo se denomina una medida descriptiva que se calcula sobre los datos de toda una población a través de un censo.

2. El puntaje promedio de ansiedad frente a los exámenes se conoce a partir de un estudio realizado de manera censal sobre todos los alumnos de una facultad y resulta ser de 75,4.
a. Cuál de los siguientes valores promedio (o valores más extremos que él) resulta menos probable de hallar en una muestra de 100 casos:
76 88 50 74,5
b. Si la muestra es de 1000 casos, la probabilidad de hallar un promedio de 90 o más extremo es, comparada con el caso de la muestra de 100:
menor mayor igual no puede saberse

4. ¿Qué significa que un estimador sea insesgado?

5. Dado que el estimador depende de la muestra y ésta depende del azar, los estimadores provenientes de muestras probabilísticas son…

6. El error estándar del estimador mide la variabilidad de…

7. ¿De qué elementos depende el error estándar del estimador?

8. Si de una población se extraen muestras y se promedian los valores de una variable, ¿qué sucede con los promedios muestrales a medida que se toman muestras de mayor tamaño?

10. El Teorema central del límite dice que una suma de variables aleatorias tiende a tener distribución normal a medida que el número de observaciones…

Capítulo 9: Estimación de parámetros

Eduardo Bologna

Hemos llegado a este punto en el que haremos uso de casi todos los elementos que se presentaron hasta aquí. Repasaremos brevemente el camino recorrido, porque desemboca de manera casi evidente en lo que aquí desarrollaremos.
Dedicamos cinco capítulos a la descripción de datos provenientes de una muestra. Luego ingresamos al terreno de la incertidumbre con el capítulo de probabilidad y continuamos con el modo de extraer una muestra representativa de una población. Finalmente el capítulo anterior a éste mostró cómo se relacionan la media de una variable cuantitativa y la proporción de una categoría de una variable cualitativa —ambas calculadas en una muestra aleatoria— con su correspondiente valor poblacional. Solo nos queda integrar estos elementos en un procedimiento para realizar las estimaciones que nos interesan. Por esta razón se trata de un capítulo de plena aplicación práctica.

Estimación puntual

La media muestral $\bar{x}$ es un estimador de la media poblacional, por lo que ya tenemos una primera estimación de ese parámetro μ. De igual modo, la proporción muestral ($\hat{p}$) estima a la proporción poblacional (P). De modo que ya tenemos estimaciones de esos dos parámetros, son estimaciones puntuales. Se llaman así porque ofrecen un único valor como estimación del parámetro de interés. Por ejemplo si en una muestra de 50 psicólogos que egresaron en los últimos diez años hallamos que han terminado la carrera con una nota promedio de $\bar{x} = 6{,}50$, disponemos de una media muestral; si ahora preguntamos por el promedio con que terminaron la carrera todos los psicólogos que egresaron en los últimos diez años, la respuesta es tentativa, diremos que "debe ser cercano a 6,50". Con esta expresión imprecisa, hacemos una estimación de la media poblacional (μ). De igual modo si en la misma muestra de 50 psicólogos, se ve que la proporción de mujeres es $\hat{p} = 0{,}70$, podremos decir que, del total de psicólogos egresados en los últimos diez años, "alrededor del 70% son mujeres". Así hacemos una estimación de P a partir de $\hat{p}$. Pero estas

estimaciones son deficientes, ya que no sabemos cuán cerca puede estar la verdadera nota promedio de 6,50 ó la verdadera proporción de mujeres del 70%. Estas son las que se denominan estimaciones puntuales.

Estimación por intervalo

Una estimación más completa de los parámetros que nos interesan, se denomina estimación por intervalo. Ella consiste en ofrecer no ya un número como en la estimación puntual, sino un intervalo, acerca del cual tendremos cierta certeza (o confianza) que contenga al parámetro. Así, en lugar de decir que el promedio con que egresa el total de psicólogos de esta facultad "debe ser cercano a 6,50", construiremos un intervalo, que dirá, por ejemplo, "tenemos una certeza del 95% que el intervalo 6,10; 6,90 contiene al promedio con que egresan los psicólogos de esta facultad". De manera equivalente, en lugar de entre los que egresan hay "alrededor del 70% de mujeres", diremos, algo como "con una certeza del 95%, el intervalo 68; 72% contiene a la proporción de mujeres sobre el total de egresados". Vemos entonces que esta forma de estimar ofrece dos números, los límites de un intervalo, del que esperamos contenga al parámetro que estimamos. Decimos "esperamos que se contenga" porque no hay certeza absoluta de que se encuentre allí, hay una confianza que en estos ejemplos hemos fijado en el 95%, y veremos que puede elegirse.

Veamos a continuación cómo construir estos intervalos de confianza para estimar los dos parámetros que venimos tratando, la media y la proporción.

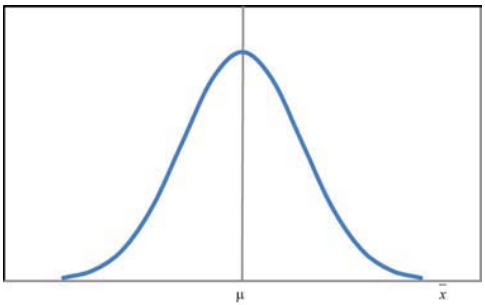
Estimación de la media

Vamos a hacer uso de lo que sabemos hasta el momento sobre las distribuciones en el muestreo para mejorar la calidad de las estimaciones puntuales y construir los intervalos de confianza. Para ello, empezaremos con la media. En el capítulo 8 dijimos que, porque la muestra ha sido sacada de manera aleatoria, la media muestral es una variable aleatoria, cuya distribución tiene media μ y desviación estándar $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Además, a medida que aumenta el tamaño de la muestra, esa distribución tiende a ser normal, es decir que será tanto más cercana a una distribución normal cuanto más grande sea n. A los fines prácticos, una muestra de 30 casos se considera "suficientemente grande" como para usar la distribución normal en la distribución de $\bar{x}$. Si la muestra es más pequeña que ese tamaño, no podemos usar inmediatamente la distribución normal, sino que deberemos apelar a la distribución t de

Student. Trabajaremos primero suponiendo que se trata de muestras lo suficientemente grandes y usaremos la distribución normal.

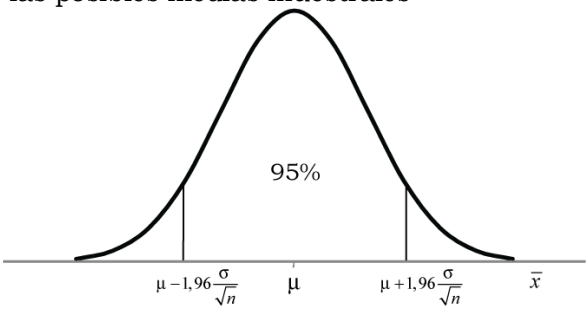
Con esa información, podemos calcular las probabilidades de los diferentes valores de $\bar{x}$. Representamos gráficamente esta distribución, como vimos antes:

Gráfico 1: Distribución de las medias muestrales



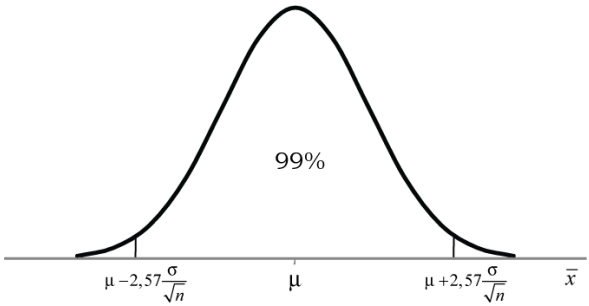
De lo que se concluye que, si extrajéramos todas las muestras de tamaño n posibles de esa población, el 95% de ellas estaría entre $\mu - 1{,}96 * \sigma_{\bar{x}}$ y $\mu + 1{,}96 * \sigma_{\bar{x}}$, o lo que es lo mismo, entre $\mu - 1{,}96 * \frac{\sigma}{\sqrt{n}}$ y $\mu + 1{,}96 * \frac{\sigma}{\sqrt{n}}$.

Gráfico 2: Intervalo en torno a la media poblacional que incluye al 95% de las posibles medias muestrales

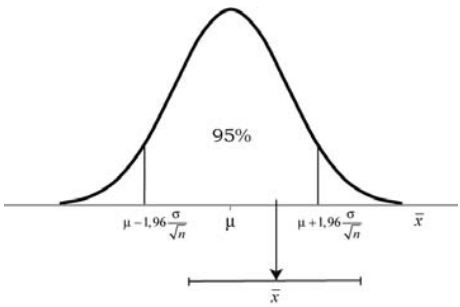


Del mismo modo, el 99% de las medias muestrales estará entre $\mu - 2{,}57 * \frac{\sigma}{\sqrt{n}}$ y $\mu + 2{,}57 * \frac{\sigma}{\sqrt{n}}$.

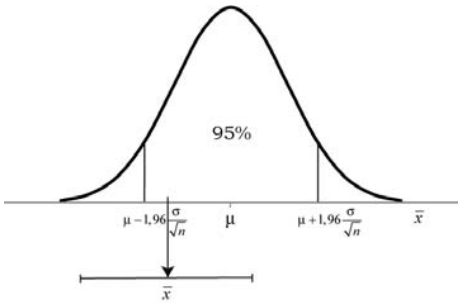
Gráfico 3: Intervalo en torno a la media poblacional que incluye al 99% de las posibles medias muestrales



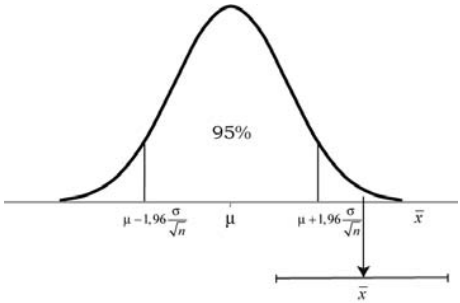
Hasta este punto lo teórico, porque las curvas de arriba solo pueden dibujarse si se conoce μ y es justamente el valor que queremos estimar. Además no se extraen "todas las muestras" sino solo una, y ella se usa para hacer la estimación. Lo que sabemos de esa muestra es que tiene una probabilidad del 0,95 de estar en la zona marcada en el gráfico 2 y una probabilidad 0,99 de estar donde indica el gráfico 3. Concentremos nuestra atención en el caso del gráfico 2, correspondiente a la zona donde se halla el 95% de todas las medias muestrales posibles.
Ahora vamos a la población y de allí sacamos una muestra (probabilística, con todos los resguardos que indicamos en el capítulo 7), en esa muestra calculamos $\bar{x}$. Supongamos que la muestra da lugar a la media que está indicada en este gráfico.



Si construimos un intervalo de la misma amplitud que el anterior, pero ahora centrado en $\bar{x}$, en vez de centrado en μ, vemos que ese intervalo contiene a μ. Si la $\bar{x}$ fuera la que está en el gráfico siguiente.



También un intervalo alrededor de ella contendría a μ. Por el contrario en el caso siguiente:



El intervalo alrededor de $\bar{x}$ no contiene a la media poblacional.

Preguntamos ¿qué condición debe cumplir $\bar{x}$ para que el intervalo que construyamos a su alrededor contenga a μ? La respuesta es que debe estar entre $\mu - 1{,}96 * \frac{\sigma}{\sqrt{n}}$ y $\mu + 1{,}96 * \frac{\sigma}{\sqrt{n}}$
¿Qué proporción de las $\bar{x}$ cumple esa condición? El 95% de ellas. Así, el 95% de las $\bar{x}$ posibles dará lugar a un intervalo que contenga a μ, el 5% restante de las $\bar{x}$ producirá intervalos que no contienen a μ.
Es importante señalar que no sabemos si nuestro intervalo contiene a μ o no, solo sabemos que hay una probabilidad de 0,95 que la contenga. Es decir, es muy probable que el intervalo contenga a μ, pero no es seguro.
¿Cuál es la expresión de ese intervalo?, dado que está centrado en $\bar{x}$, hay que sumar y restar a ese estimador lo mismo que sumamos y restamos a μ para construir el intervalo anterior, por lo que resulta:

$$\bar{x} - 1{,}96 * \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + 1{,}96 * \frac{\sigma}{\sqrt{n}}$$

Estos son dos números que constituyen los límites de un intervalo que tiene una probabilidad 0,95 de contener al parámetro μ. De manera equivalente decimos que, de cada 100

intervalos que se construyan con este procedimiento, 95 contendrán a la media de la población. O bien que el 95% de las muestras aleatorias de tamaño n que se extraigan de la población, proveerán valores de $\bar{x}$ que conducirán a intervalos que contengan a la media de la población.
Cuando logramos construir un intervalo así decimos que estimamos a μ con un 95% de confianza. El primero valor de los indicados se llama límite inferior (L_i) y el segundo, límite superior (L_s). Así entonces:

$$Li = \bar{x} - 1{,}96 * \frac{\sigma}{\sqrt{n}}$$
$$Ls = \bar{x} + 1{,}96 * \frac{\sigma}{\sqrt{n}}$$

Veamos un ejemplo. Si en una muestra de 400 egresados de Psicología, encontramos que la nota promedio con que egresan es de 6,50 ($\bar{x} = 6{,}50$) y sabemos que la desviación estándar de la población es de 0,8 ($\sigma = 0{,}8$), estimamos la nota promedio con que egresaron todos los psicólogos reemplazando:

$$Li = 6{,}50 - 1{,}96 * \frac{0{,}8}{\sqrt{400}} = 6{,}42$$
$$Ls = 6{,}50 + 1{,}96 * \frac{0{,}8}{\sqrt{400}} = 6{,}58$$

Leemos este resultado diciendo que tenemos un confianza del 95% que el intervalo (6,42; 6,58) contiene a la media de las notas con que egresaron todos los psicólogos de esta facultad. La confianza del 95% está incluida en la construcción del intervalo en el número 1,96 que multiplica al error estándar de $\bar{x}$.

La notación puede abreviarse indicando de una sola vez ambos límites, si se escribe:

$$\bar{x} \pm 1{,}96 * \frac{\sigma}{\sqrt{n}}$$

Con lo que queremos indicar que a $\bar{x}$ primero le sumamos y luego le restamos la expresión $1{,}96 * \frac{\sigma}{\sqrt{n}}$.
En el ejemplo, esto haría que escribamos el intervalo de manera alternativa como $6{,}50 \pm 0{,}08$, que indica cuál es la media muestral (el estimador puntual) y la cantidad que debe sumarse y restarse para llegar a los límites.

Si quisiéramos estar más seguros acerca de que el intervalo contiene a μ, podríamos usar los puntos que delimitan el 99%

del área. Para ello, z vale 2,57 y los límites del intervalo resultan: $\bar{x} \pm 2{,}57 * \frac{\sigma}{\sqrt{n}}$
Para el ejemplo anterior, con una confianza del 99%, el intervalo es:

$$L_i = 6{,}50 - 2{,}57 * 0{,}04 = 6{,}50 - 0{,}10 = 6{,}40$$
$$L_s = 6{,}50 + 2{,}57 * 0{,}04 = 6{,}50 + 0{,}10 = 6{,}60$$

Con lo que ahora diremos que, con una confianza del 99%, el intervalo (6,40; 6,60) contiene a la media de las notas con que egresaron todos los psicólogos de esta facultad. Otra opción es la de escribir el intervalo como $6{,}50 \pm 0{,}10$, la media muestral es la misma y aumentó lo que debe alejarse de ella para llegar a los límites.

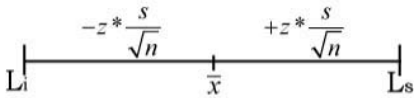
Notemos que este aumento en la confianza de la estimación, al pasar del 95 al 99%, tiene un costo, porque el intervalo es ahora más amplio: el límite inferior es menor que en el anterior y el superior, mayor. Antes el intervalo iba desde 6,42 a 6,58 y ahora va desde 6,40 hasta 6,60. Más tarde volveremos sobre este punto.

De manera general, escribiremos el intervalo como $\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$, dejando z como variable, que puede reemplazarse por el valor que corresponda según la confianza que se elija para la estimación.
Sin embargo, esta manera de calcular los límites del intervalo tiene un problema para usarse en la práctica, ya las fórmulas para calcular los límites, requieren que se conozca σ, la desviación estándar de la población. Pero como nuestros datos son muestrales, no conocemos σ, a cambio de ella usaremos la desviación estándar de la muestra, a la que podemos calcular a partir de los datos disponibles[68]. Con ese ajuste, la expresión para el cálculo de los límites del intervalo de confianza será:

$$\bar{x} \pm z * \frac{s}{\sqrt{n}}$$

[68] Esto es válido en la medida que se trate de muestras grandes (*n>30*), en caso contrario, la distribución que debemos usar es la *t* de Student. Cuando fijemos la confianza, ya no serán *z* los valores que multiplicarán a $\frac{\sigma}{\sqrt{n}}$ sino puntajes *t*, cuyos grados de libertad se calculan como *n-1*. Pero, para poder usar la distribución *t*, los valores de la muestra deben provenir de una distribución normal en la población. Si esto no se cumple, la estimación solo será aproximada.

Que puede representarse gráficamente así:



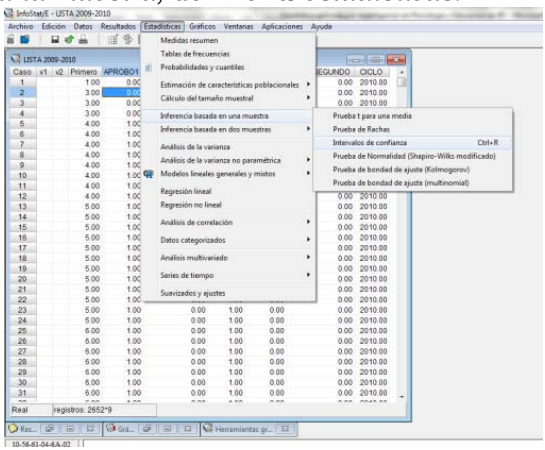
En este gráfico solo podemos dibujar el segmento que representa al intervalo en torno a $\bar{x}$, pero no podemos dibujar la campana correspondiente a la distribución, ya que no conocemos μ que es donde la campana se centra.

Veamos un ejemplo con datos reales. Disponemos de una muestra de 277 alumnos que rindieron el primer parcial, conocemos sus notas y queremos usarlas para hacer una estimación de la nota promedio de todo el curso (que cuenta con 1600 alumnos). Haremos esa estimación con una confianza del 95%.
De la muestra hemos obtenido $\bar{x} = 6{,}76\ y\ s = 2{,}13$, con lo que los límites resultan:

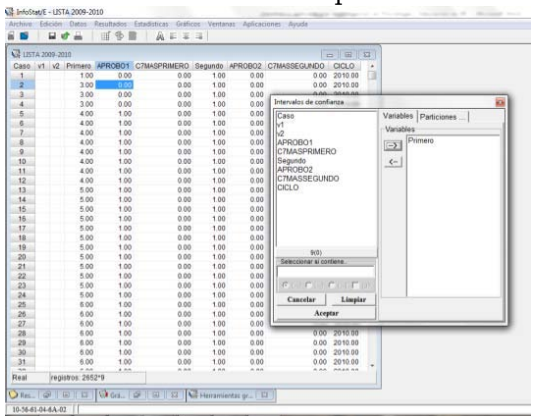
$$\bar{x} \pm z * \frac{s}{\sqrt{n}} = 6{,}76 \pm 1{,}96 * \frac{2{,}13}{\sqrt{277}} = 6{,}76 \pm 1{,}96 * 0{,}13 = 6{,}76 \pm 0{,}25$$

Usando primero el signo menos, obtenemos L_i = 6,51 y luego sumando L_s = 7,01. Entonces podemos afirmar el intervalo 6,51; 7,01 contiene a la nota promedio del total de alumnos del curso, con una confianza del 95%.

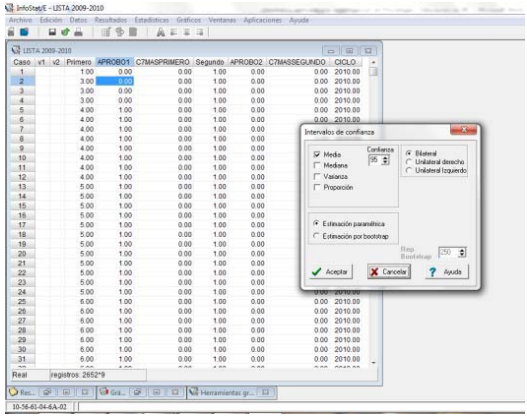
Para pedir esta operación a InfoStat®, lo hacemos desde *intervalos de confianza*, que está entre las opciones de *inferencia basada en una muestra*, del menú *estadísticas*:



Una vez que introducimos la variable primero



Disponemos de algunas opciones para el intervalo



Como estamos estimando la nota promedio, dejamos marcado media. Luego seleccionamos el nivel de confianza, 95%, y obtenemos:

Intervalos de confianza

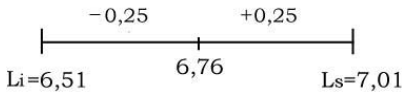
Bilateral

Estimación paramétrica

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Primero	Media	6,76	0,13	277	6,51	7,01

La variable se llama "primero" en referencia al primer parcial. El E.E. es el error estándar del estimador, proviene de $\frac{s}{\sqrt{n}}$, que es $\frac{2,13}{\sqrt{277}} = 0,13$. Ese error estándar del estimador es el que se multiplica por z para obtener el término que conduce a los límites del intervalo.

Una representación gráfica de esta estimación es:

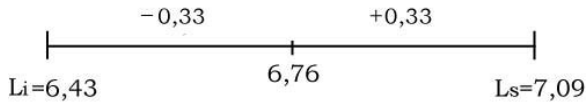


Si la confianza se solicita en el 99%, el programa da:

Intervalos de confianza
Bilateral
Estimación paramétrica

Variable	Parámetro	Estimación	E.E.	n	LI(99%)	LS(99%)
Primero	Media	6,76	0,13	277	6,43	7,09

Comparada con la salida anterior, solo han cambiado los límites del intervalo, ya que son los mismos datos muestrales. El cambio en la confianza se realiza por un cambio en el valor de z y eso hace que cambien los límites. El gráfico tiene ahora la forma:



Como ya habíamos visto, un aumento en la confianza incide en la amplitud del intervalo, éste último es más amplio que el primero que teníamos. Más adelante trataremos esta relación con detalle.

La estimación de la proporción

Cuando trabajamos con variables cualitativas (nominales u ordinales) no es posible calcular la media ni la desviación estándar sino meramente considerar la proporción de casos que hay en una categoría que elegimos. Cuando se trata de variables con solo dos categorías (dicotómicas) puede elegirse cualquiera de ellas. Por ejemplo si trabajamos con el resultado de un examen y las categorías son aprobado – no aprobado, podemos interesarnos por la proporción de cualquiera de ellas, ya que la otra es el complemento (lo que le falta para llegar a uno). Si una es 0,70, la otra no puede sino ser 0,30. Es diferente si la variable tiene más de dos categorías, por ejemplo si se trata de la intención de voto para las elecciones presidenciales. Allí es usual que haya más de dos candidatos, por lo que la proporción de uno de ellos no nos dice mucho sobre la de cada uno de los otros: si hay cinco candidatos y uno se lleva el 40%, solo

sabemos que el 60% restante se reparte entre los otros cuatro, pero no sabemos cuánto le corresponde a cada uno. A estos casos los trataremos como si fueran dicotómicos: una categoría será el candidato que nos interesa y la otra categoría estará formada por todos los demás. Así, si un candidato tiene una proporción de 0,40 a su favor, solo nos interesa que tiene una proporción de 0,60 que no está a su favor y no nos preocupamos por saber cómo se reparte ese 60% en los demás candidatos. Tratamos una categoría frente a todas las demás. De este modo es que puede definirse la proporción de personas que usa anticonceptivos orales, frente a quienes usan todos los demás métodos; o la proporción de alumnos promocionados frente a regulares y libres; o la proporción de argentinos entre todos los estudiantes extranjeros que hay en España, sin interesarnos por el modo en que se distribuye la proporción entre las demás nacionalidades. Lo que hacemos con este procedimiento es simplemente llamar la atención sobre una categoría y confrontarla con el resto indiscriminado.
Por este procedimiento trataremos siempre con dos grupos, uno formado por los casos que son de nuestro interés y el otro por los demás casos.

El razonamiento que seguimos para la estimación de P es análogo al que seguimos para estimar μ. La estructura de los límites del intervalo de confianza es ahora:

$$L_i = \hat{p} - z * \sigma_{\hat{p}}$$
$$L_s = \hat{p} + z * \sigma_{\hat{p}}$$

En la que:
$\hat{p}$ es la proporción de casos en la categoría que estimamos calculada sobre los datos de la muestra.
z asume el valor de 1,96 si vamos a estimar con una confianza del 95%, ó de 2,57 si queremos una confianza del 99%.
$\sigma_{\hat{p}}$ nos es conocida desde el capítulo anterior, y vale:

$$\sigma_{\hat{p}} = \sqrt{\frac{P * (1 - P)}{n}}$$

Pero, tal como pasó con la estimación de $\bar{x}$, en la que ignorábamos σ por tratarse de un valor poblacional, ahora desconocemos P (¡es exactamente lo que estamos tratando de estimar!), por lo que deberemos necesariamente reemplazarla por su estimador: $\hat{p}$[69]. Nos quedará: $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$
Y los límites del intervalo resultarán:

[69] Como antes hicimos reemplazando a σ por s.

$$\hat{p} \pm z \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

Lo aplicamos a un ejemplo: se trata de la muestra de alumnos que rindieron el parcial a partir de la cual queremos estimar, al 95%, la proporción de quienes lo aprobaron. Sabemos que, de los 277 que rindieron, 255 lo aprobaron, en consecuencia la proporción de aprobados es $\hat{p} = \frac{255}{277} = 0{,}920$. Este es nuestro estimador puntual de la proporción de aprobados para todo el curso. Para hacer el intervalo, usamos la expresión anterior y resulta:

$$\hat{p} \pm z \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}} = 0{,}920 \pm 1{,}96 * \sqrt{\frac{0{,}92 * 0{,}08}{277}} = 0{,}920 \pm 0{,}032$$

Cuando restamos, obtenemos el límite inferior del intervalo:

$$L_i = 0{,}920 - 0{,}032 = 0{,}892$$

y sumando:

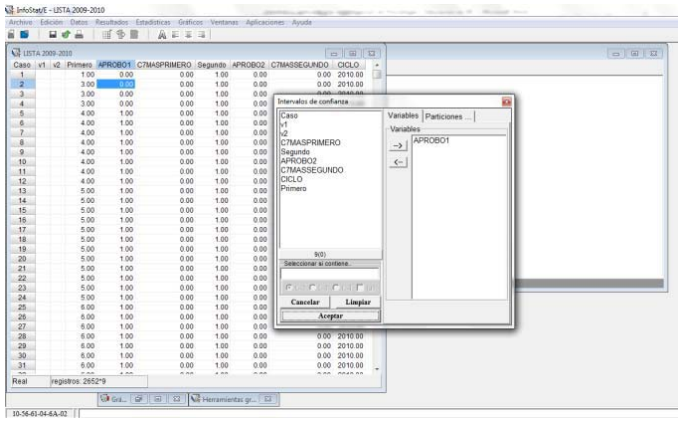
$$L_s = 0{,}920 + 0{,}032 = 0{,}952$$

Si se escribe de manera abreviada, la expresión toma la forma:

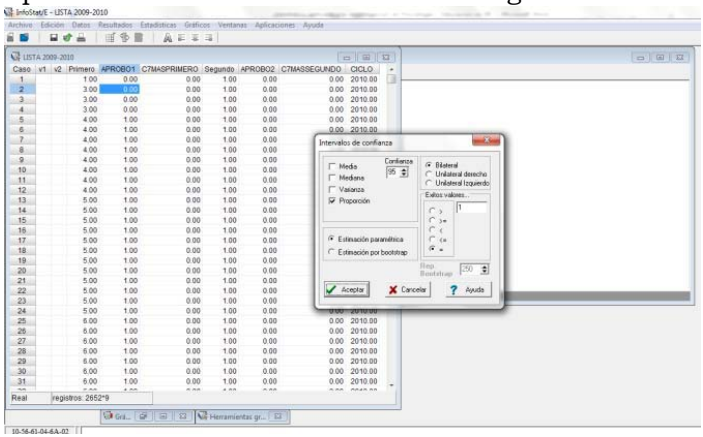
$$L_i = 0{,}920 \pm 0{,}032$$

Con el valor explícito de la proporción muestral que es el estimador puntual de *P*.
El resultado nos dice que hay una certeza del 95% que el intervalo 0,892; 0,952 contenga a la proporción de aprobados de toda la población.

Para solicitar la operación a InfoStat®, se selecciona la variable que indica el resultado del parcial, en esta matriz de datos se llama *aprobó1*:



En la ventana siguiente debemos especificar que no estimamos la media, sino la proporción. Además hay que indicar cuál es la categoría que se considera "éxito", es decir, cuál es la categoría cuya proporción nos interesa. En este ejemplo, la variable *aprobó1* está codificada como:
1 = aprobó
0 = no aprobó
Por lo que indicamos como éxito los valores iguales a 1:



La salida tiene forma:

Intervalos de confianza

Bilateral

Estimación paramétrica

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Primero	Proporción(=1)	0,92	0,02	277	0,89	0,95

La indicación entre paréntesis (=1) se refiere a la condición que pusimos como aprobado.

Si aumentamos la confianza al 99%, el intervalo es:

Intervalos de confianza
Bilateral
Estimación paramétrica

Variable	Parámetro	Estimación	E.E.	n	LI(99%)	LS(99%)
Primero	Proporción(=1)	0,92	0,02	277	0,88	0,96

Como sucedió con la media, el aumento en la confianza se traduce en un aumento en la amplitud del intervalo.

La calidad de las estimaciones por intervalo

Intuitivamente, una estimación es de mejor calidad si es "ajustada", es decir si el intervalo es pequeño. Por ejemplo, si estimamos la edad de una persona entre 28 y 30 años, tenemos una estimación de mejor calidad que si decimos que tiene entre 20 y 40 años. Eso es porque el primer intervalo es más pequeño, los límites están más cerca. La primera estimación nos da más información que la segunda, porque delimita el valor al que estima entre números más cercanos. En las estimaciones que hemos hecho hasta aquí, de la media y de la proporción, hemos partido del estimador puntual ($\overline{x}$ y $\hat{p}$) y desde él sumamos y restamos la misma cantidad para obtener los límites del intervalo.
Esa cantidad que sumamos y restamos determina la amplitud del intervalo: cuanto más grande sea, tanto mayor será el intervalo, tanto mayor será la distancia entre los límites inferior y superior. Esa cantidad se denomina **error de estimación**. Es la distancia que hay desde el centro del intervalo hasta cualquiera de los límites. En el ejemplo anterior, sobre la estimación intuitiva de la edad de alguien, el centro del primer intervalo es 29, por lo que el error es 1 año, por eso se puede también escribir como 29 ± 1. El segundo intervalo de este ejemplo tiene centro en 30 y el error es de 10 años, lo escribimos 30 ± 10. Independientemente que el centro de los intervalos difiera levemente, este segundo intervalo tiene un mayor error de estimación. Esto es equivalente a decir que tiene menos **precisión**.

Se llama **error de estimación** a la distancia que hay entre el estimador puntual y cualquiera de los límites del intervalo.
Cuanto mayor es el error de estimación menor es su **precisión**.

En la estimación del promedio con que egresan los psicólogos, escribimos $6{,}50 \pm 0{,}08$ al estimar al 95% de confianza y $6{,}50 \pm 0{,}10$ cuando la confianza se pasó al 99%. Allí estábamos escribiendo el intervalo como el estimador más/menos el error de estimación. En el primer caso el error de estimación es de 0,08 y en el segundo de 0,10, por eso decimos que la primera estimación es más precisa.
Del mismo modo, al estimar, con una confianza del 95%, la proporción de quienes aprobaron el parcial escribimos $0{,}920 \pm 0{,}032$, el error de estimación es en este caso de 0,032 (ó 3,2%).

El error de estimación en la media
En la expresión general de la estimación por intervalo de μ, el error es el término que se suma y resta: $z * \frac{s}{\sqrt{n}}$. ¿De qué depende que ese término sea grande o chico?

Hay tres elementos en este término: *z*, *s* y *n*. De ellos va a depender que haya más o menos error en la estimación o, dicho de otra manera, que la estimación sea más o menos precisa. Veamos el efecto de cada uno:

z: Es elegido por el investigador cuando establece la confianza. En los ejemplos que hemos visto, asumió el valor de 1,96 para un 95% de confianza ó de 2,57 para una confianza de 99%. Cuanto más confianza o certeza queramos tener en nuestra estimación, más grande será *z* y, en consecuencia mayor será el error de estimación. Por lo tanto no se pueden tener las dos cosas: más confianza va acompañada de menos precisión. Si todos los demás elementos del error quedan fijos, los intervalos más amplios proveen menos información, pero mayor certeza en la inclusión del parámetro que se estima. Para elegir el nivel de confianza (y en consecuencia determinar *z*) debe tomarse una decisión que equilibre la confianza y la precisión, ya que si una crece la otra disminuye.

s: La desviación estándar en la muestra. Es la medida de la variabilidad de los datos que observamos y es una estimación de la verdadera variabilidad que tiene la característica que estamos estudiando, en la población. Incide negativamente sobre el error, cuanto más grande es *s* más error tenemos. Eso refleja el hecho que si la población es muy heterogénea respecto de la cualidad que queremos estimar, tendremos estimaciones de peor calidad que si es similar para los individuos de la población. Sobre *s* no podemos decidir, no tenemos control sobre su valor, si es

grande, tendremos peores estimaciones que si es pequeña. El muestreo estratificado es una forma de enfrentar situaciones de mucha dispersión, construyendo subconjuntos (estratos) que contengan elementos homogéneos en su interior, es decir que tengan menos dispersión que el conjunto completo.

n: El tamaño de la muestra, se encuentra en el denominador del término del error, por lo que su aumento reduce el error. Cuanto más grande sea *n*, menor será el error, es decir que muestras de mayor tamaño dan mayor precisión. En principio, podemos elegir *n*, pero depende del presupuesto que se prevea para la investigación. Si se puede obtener una muestra grande siempre es preferible, porque se lograrán estimaciones de mejor calidad.

Esto no debe confundirse con la calidad de la muestra. Todo lo que hemos dicho en el capítulo 8 y en este, supone que se trata de muestras probabilísticas, es decir muestras aleatorias, para las cuales rigen las leyes de probabilidad que hemos usado. Si la muestra no es aleatoria, no se pueden hacer estimaciones con estos procedimientos y, es muy importante; no se mejora una muestra tomando más casos. Si la muestra no es probabilística, la estimación no mejorará porque se tomen muchos casos.

Para ejemplificar los efectos de los diferentes elementos en el error de estimación, volveremos sobre los datos de las notas del primer parcial y haremos tres diferentes estimaciones de la nota promedio:

Sobre el total de la muestra (277 casos) al 90% de confianza:

Variable	Parámetro	Estimación	E.E.	n	LI(90%)	LS(90%)
Primero	Media	6,76	0,13	277	6,55	6,97

Sobre el total de la muestra (277 casos) al 95% de confianza:

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Primero	Media	6,76	0,13	277	6,51	7,01

Sobre el total de la muestra (277 casos) al 99% de confianza:

Variable	Parámetro	Estimación	E.E.	n	LI(99%)	LS(99%)
Primero	Media	6,76	0,13	277	6,43	7,09

Estas tres primeras estimaciones muestran cómo, sin cambiar el tamaño de la muestra ni la dispersión, el error aumenta (los intervalos se vuelven más amplios) cuando crece la confianza.

Sólo sobre el turno tarde (85 alumnos) al 95% de confianza:

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Primero	Media	6,78	0,25	85	6,28	7,27

Si comparamos esta estimación con la segunda, que está hecha también al 95%, vemos que es menos precisa, ya que la distancia entre los límites es de $7{,}01 - 6{,}51 = 0{,}50$ en el primer caso y $7{,}27 - 6{,}28 = 0{,}99$. Esto se debe al menor tamaño de muestra: menos casos, más error de estimación.

El error de estimación en la proporción

El término del error en la estimación de la proporción es

$$z * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

En él hay dos elementos en común con el error en la estimación de la media: los valores de z y de n. No agregaremos nada sobre ellos, porque el efecto es el mismo que en la media: un aumento de z por aumento de la confianza, incrementa el error de estimación; un aumento en el tamaño de la muestra, lo reduce.

Lo nuevo en este caso es que no hay s, por el contrario, lo que hay en su lugar es el producto de la proporción por su complemento $\hat{p} * (1 - \hat{p})$, que se encuentra afectado por la raíz, pero eso no nos va a interesar para analizar su efecto sobre la precisión.

Recordemos el problema de la medición de la dispersión para variables nominales que tratamos en el capítulo 3 y retomamos en el 8. Allí dijimos que una variable nominal tiene poca dispersión cuando una categoría "absorbe" a las otras, cuando muchos casos están en una sola categoría, o cuando una categoría tiene una frecuencia superior a todas las demás. Por el contrario, la dispersión es elevada cuando las frecuencias son similares, cuando la distribución de casos es "pareja" en todas las categorías. En la estimación de la proporción estamos tratando solo con dos categorías, por lo que la dispersión será máxima cuando las proporciones de ellas sean similares. Siendo solo dos, son iguales cuando cada una de ellas vale 0,50 ($\hat{p} = 0{,}50\ y\ (1 - \hat{p}) = 0{,}50$), porque la mitad de los casos está en cada categoría. Por el contrario, la dispersión será menor cuanto más concentrados estén los casos en una de las categorías. Si, por ejemplo la proporción es 0,10 ($\hat{p} = 0{,}10\ y\ (1 - \hat{p}) = 0{,}90$) tendremos concentración de casos en una categoría, es decir,

poca dispersión. Eso está expresado en la variabilidad medida como el producto de $\hat{p}$ por su complemento: $\hat{p} * (1 - \hat{p})$.
Cuando $\hat{p} = 0{,}50\ y\ (1 - \hat{p}) = 0{,}50$, entonces, el producto $\hat{p} * (1 - \hat{p}) = 0{,}25$. Por el contrario, cuando $\hat{p} = 0{,}10\ y\ (1 - \hat{p}) = 0{,}90$, entonces, el producto $\hat{p} * (1 - \hat{p}) = 0{,}09$.
Por eso, el producto $\hat{p} * (1 - \hat{p})$ es una medida de la dispersión de la variable nominal y ocupa, dentro del término del error, un lugar equivalente al de la varianza en la estimación de la media.
¿Cómo incide esto en el error de estimación? Como con la media, cuando la dispersión es grande, el error también lo es, entonces el error será mayor cuanto más parecidas sean $\hat{p}\ y\ (1 - \hat{p})$, dicho de otra manera, cuando $\hat{p}$ sea cercana a 0,50.
El razonamiento es el mismo que con la media, cuanto mayor sea la dispersión tanto más grande será el error y menos precisa la estimación. Pero en el caso de la media, la dispersión está medida con la desviación estándar, mientras que en la proporción, viene dada por el producto $\hat{p} * (1 - \hat{p})$, que es máximo cuando $\hat{p}$ es cercano a 0,50. Entonces, las peores condiciones para hacer una estimación de la proporción, serán aquellas en que la característica que se estima afecta a porciones cercanas a la mitad de la muestra, allí será máxima la dispersión y en consecuencia también el error de estimación.
A partir de una encuesta, se estima la proporción de votos que tendrá un candidato en las próximas elecciones. La muestra es de 400 casos y 90 personas dijeron que votarán a ese candidato. Como 90 es el 22,5% de 400, esa es la proporción que se halla en la muestra y la estimación por intervalo al 95% de confianza nos da:

$$\hat{p} \pm z * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}} = 0{,}225 \pm 1{,}96 * \sqrt{\frac{0{,}225 * (1 - 0{,}225)}{400}}$$
$$= 0{,}225 \pm 0{,}041$$

Y los límites del intervalo son L_i = 0,1841 y L_s = 0,2659. Para comunicarlo, diremos que el candidato tiene una intención de voto de entre el 18,41% y el 26,59%.

Repitamos el ejercicio, ahora suponiendo que la cantidad de personas que dice que lo votaría son 200 de los 400 encuestados, es decir si la proporción muestral hubiese sido del 50%. Siempre al 95% de confianza, la estimación es:

$$\hat{p} \pm z * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}} = 0{,}50 \pm 1{,}96 * \sqrt{\frac{0{,}50 * (1 - 0{,}50)}{400}} = 0{,}50 \pm 0{,}049$$

Vemos que el error de estimación ha pasado de 4,1% en el anterior a 4,9% ahora, sin que hayamos cambiado la confianza

ni el tamaño de la muestra. Ese es el efecto de la proporción cuando es cercana al 50%.

Solicitamos nuevamente a InfoStat® la estimación de la proporción de aprobados en diferentes condiciones:

Sobre el total de la muestra (277 casos) al 90%

Variable	Parámetro	Estimación	E.E.	n	LI(90%)	LS(90%)
Primero	Proporción(=1)	0,921	0,016	277	0,894	0,947

Sobre el total de la muestra (277 casos) al 95%

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Primero	Proporción(=1)	0,921	0,016	277	0,889	0,952

Sobre el total de la muestra (277 casos) al 99%

Variable	Parámetro	Estimación	E.E.	n	LI(99%)	LS(99%)
Primero	Proporción(=1)	0,921	0,016	277	0,879	0,962

Vemos que al aumentar el nivel de confianza se reduce la precisión, ya que los límites se distancian, volviendo más amplio al intervalo.

Solo con el turno noche (110 alumnos) al 95%

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Primero	Proporción(=1)	0,918	0,026	110	0,867	0,969

El intervalo se amplía respecto de la segunda de las estimaciones anteriores, aunque la confianza es la misma (95%), debido a la reducción en el número de casos.

En este capítulo hemos puesto en juego lo visto en los anteriores para poder generalizar las observaciones muestrales a toda la población de referencia, vemos que el modo con el que se hace es a través de los intervalos de confianza, que formalizan una práctica a la que estamos acostumbrados cuando hacemos estimaciones sobre cantidades que desconocemos: indicamos entre qué valores es más probable hallarlas.
La estructura general de los intervalos es:

$$estimador \pm z * error\ estandar\ del\ estimador$$

Esa expresión ha tomado dos formas, ya sea para estimar la media de variables cuantitativas o la proporción de casos en una categoría, cuando se trata de variables cualitativas.

La lectura del intervalo obtenido se expresa:

Hay una confianza $1 - \alpha$ que el intervalo obtenido contenga al parámetro.

Actividad práctica de repaso 9

1. Se estima la edad promedio de la población de estudiantes universitarios a partir de una muestra de 200 casos. A un nivel del 95%, se obtiene el siguiente intervalo: 22,5; 27,5.
a. ¿Cuánto vale el estimador puntual?
b. ¿Cuál es el error de estimación?
c. ¿Cuál es el límite inferior del intervalo de confianza?
d. Redacte una lectura del intervalo.
e. Si la confianza se reduce al 90%, ¿qué sucede con el error de estimación?

2. Se estima el tiempo requerido para responder a un cuestionario, usando datos de una muestra de 100 aplicaciones y —al 90%—, se obtiene 5±1 minutos.
a. ¿Cuál es el estimador puntual?
b. ¿Cuánto vale el error de estimación?
c. ¿Cuál es el límite superior del intervalo de confianza?
d. Realice una lectura del intervalo.

3. En una prueba de atención se consideran sobresalientes a quienes cometen menos de dos errores. Cuando se aplica a una muestra de 300 pacientes diagnosticados de depresión, se observa que 51 de ellos alcanzan el nivel "sobresaliente".
a. ¿Cuál es la proporción de quienes cometieron menos de dos errores en la muestra?
b. ¿Cuánto vale el error estándar del estimador?
c. ¿Cuáles son los límites del intervalo con una confianza de 95%?
d. Redacte una lectura del intervalo
e. Si la confianza se aumenta al 99%, se espera que:

4. En una muestra de personas adultas, conductoras de vehículos particulares, el 20% dice que no es grave pasar un semáforo en rojo si no viene nadie por la otra calle. Cuando se expande a la población, con una confianza del 95% se encuentra el intervalo (18; 22)%.
a. ¿Cuánto vale el estimador puntual?
b. ¿Cuánto vale el error de estimación?
c. Redacte una lectura del intervalo.
d. Un aumento en el número de casos de la muestra, sin cambiar el resto de las condiciones de la estimación, ¿qué efecto tiene?

5. Para estimar la proporción de personas para quienes tiene efectividad un determinado tratamiento, se dispone de una muestra de 500 individuos, de los cuales 400 tuvieron resultados positivos luego del tratamiento. Al 95% de confianza, el error de estimación resulta ser de 3,5%.
a. ¿Qué parámetro se estima?
b. ¿Cuánto vale el estimador puntual?
c. ¿Cuál es el límite inferior del intervalo de confianza?
d. Realice una lectura del intervalo
e. Si el número de casos fuera de 300 en lugar de 500, y todo lo demás se mantuviera sin cambios, se esperaría que el error de estimación…

Capítulo 10: Las pruebas de hipótesis

Eduardo Bologna
Cecilia Reyna

En este capítulo ingresamos plenamente a uno de los procedimientos de mayor difusión en las investigaciones en Psicología: la prueba de hipótesis; y veremos que usa los mismos principios de estimación de parámetros en que se basa la construcción de intervalos de confianza que tratamos en el capítulo anterior. Aunque no está exento de críticas, este procedimiento es básico para hacer inferencias sobre la población, y las propuestas que existen como técnicas alternativas, requieren que se tenga dominio de las pruebas de hipótesis.

El razonamiento de la prueba de hipótesis

La prueba de hipótesis tiene como objetivo el de darnos argumentos para decidir en contextos de incertidumbre. A partir de lo que sabemos sobre las distribuciones en el muestreo, ése es el caso cuando necesitamos concluir acerca de una población, a partir de información que tenemos disponible en una muestra aleatoria. El resultado de la prueba permitirá decidir si lo que se observa en la muestra es compatible con una aseveración hipotética sobre la población. Nunca será posible decidir de manera taxativa que la hipótesis es verdadera, eso es algo que no podemos saber; por el contrario, podemos ver hasta qué punto lo que observamos en la muestra contradice —o no— lo que se afirma a escala poblacional. Es decir que podremos descartar una hipótesis por no ser compatible con lo que se observa, pero no a la inversa: no será posible "confirmar" una hipótesis, solo podremos concluir que la evidencia *no la contradice*, lo que también se expresa diciendo que *no hay evidencia para rechazarla*.

Empezaremos con ejemplos no muy cercanos a la estadística, a fin de ver que esta forma de razonar no es para nada ajena a lo cotidiano. Una prueba de hipótesis puede compararse con un juicio: el acusado no es condenado hasta que no hay evidencia suficiente para hacerlo. La evidencia (las pruebas, en el lenguaje de la justicia) rara vez son completas, se trata de información

fragmentada, sujeta a interpretaciones diferentes. En el inicio del juicio, "el acusado es inocente", en nuestra notación llamaremos a esa afirmación, **hipótesis nula**, y la indicaremos **Ho** Esta expresión indica que se trata de un estado inicial: todos son inocentes hasta que se prueba lo contrario, por lo que la hipótesis nula señala que esta persona en particular (el acusado), no es diferente de cualquier ciudadano que no ha cometido delito. Mientras no haya pruebas suficientes, la hipótesis nula se considerará aceptada. En el juicio, el fiscal aportará pruebas en dirección contraria a esta hipótesis. Buscará información para probar que debe rechazarse la hipótesis nula y condenar al acusado. Difícilmente estarán a la vista todos los datos necesarios para reconstruir la situación y dar una respuesta absolutamente inequívoca, pero si hay suficiente evidencia, se dará la hipótesis nula por rechazada. La decisión de condenar al acusado solo se tomará cuando haya muy poco riesgo de equivocarse, cuando la probabilidad de decidir de manera errada sea muy pequeña.
En este ejemplo, la población es el conjunto completo de información necesaria para tomar la decisión de manera certera. Se trataría de un conjunto de datos muy amplio, que no está disponible, por lo que la decisión debe tomarse a partir de un fragmento de información, que son las pruebas que han podido reunirse, en la analogía que hacemos con nuestros procedimientos, esto constituye la muestra a partir de la que se tomará la decisión sobre la hipótesis nula: aceptarla o rechazarla.

Otro ejemplo: tenemos dudas sobre lo equilibrada que pueda estar una moneda que va a usarse en un juego de azar. Repitiendo la notación del ejemplo anterior, formularemos una hipótesis nula que dice que X sale con la misma frecuencia que C, que equivale a decir que hasta que no se pruebe lo contrario, X no tiene ninguna diferencia con C, la moneda está equilibrada. Esta hipótesis nula puede escribirse de manera formal, porque "salir con la misma frecuencia que C" equivale a decir que, "en muchas tiradas, la mitad de la veces saldrá X", por lo que escribiremos nuestra hipótesis nula así: H_0: P=1/2, a la que leeremos "la hipótesis nula afirma que la proporción de veces que saldrá cara es 1/2".
Luego debemos producir los datos para hacer la prueba, al tirar la moneda 100 veces esperaríamos —si la hipótesis nula se sostiene—, que salga aproximadamente 50 veces X. Sabemos que las 100 tiradas son una muestra de las infinitas tiradas de la moneda, por lo que posiblemente no salga exactamente 50

veces X, podría salir 51 veces ó 52 y serían resultados esperables, debido a fluctuaciones propias del azar. Pero si de las 100 tiradas sale 80 veces X, concluiremos con pocas dudas que hay que rechazar la hipótesis nula. A la misma conclusión llegaríamos si, de 100 tiradas, solo sale 25 veces X. La pregunta que nos ayudará a responder el procedimiento de prueba de hipótesis es ¿cuántas más o menos veces que 50 debería salir X para que consideremos que tenemos "suficiente evidencia", para creer que la moneda no está equilibrada?

La lógica de la prueba de hipótesis consiste en plantear el escenario en el que H_0 es verdadera y observar qué tan probable es lo que hallamos en la muestra en ese caso. En el primer ejemplo el planteo es ¿qué tan probable sería haber hallado estas pruebas contra el acusado, si éste fuera inocente? En el segundo preguntamos ¿qué tan probable habría sido hallar esta cantidad de veces que salió X, si la moneda estuviera equilibrada? De manera general la pregunta es ¿qué tan probable sería éste resultado muestral si la hipótesis nula fuera cierta?
Si la respuesta a esas preguntas es "muy probable", la decisión será la de no rechazar la hipótesis nula, porque los resultados muestrales hallados serían esperables (muy probables) bajo H_0. Al contrario, si la respuesta es "muy poco probable" decidiremos rechazar H_0, ya que se trata de un resultado poco esperable si H_0 fuera cierta.
En investigación, la prueba de hipótesis suele formularse de tal modo que rechazar H_0 implica aportar un nuevo hallazgo, por el contrario, aceptar H_0 equivale a que no hay cambios respecto de la situación inicial.
Algunos ejemplos de hipótesis nulas:
- Esta droga no produce ningún efecto sobre la memoria.
- La técnica terapéutica A es igualmente eficaz que la B.
- Los métodos A y B para enseñar a leer a los niños producen iguales resultados.
- La proporción de votos que obtendrá un candidato no ha variado respecto de las últimas elecciones.

En casi todos los casos, la expectativa del investigador está en rechazar la H_0, porque eso significa que ha hallado algo de interés: que la droga produce efectos, que hay técnicas terapéuticas mejores que otras y por tanto recomendables, que se pueden elegir mejores métodos para enseñar a leer, que el favor del electorado hacia un político es más o menos extendido.

La hipótesis nula es una afirmación sobre un parámetro que indica ausencia de diferencia.

El sentido de esa diferencia difiere según el tipo de prueba, veremos que hay pruebas que confrontan con valores históricos o con promedios generales o bien que realizan comparaciones entre grupos. En todos los casos la hipótesis nula afirma que "no hay diferencia"

Hasta este punto se trata de la definición original de estas pruebas, desarrolladas inicialmente por Sir Ronald Fisher (1925) a las que llamó *pruebas de significación*. La idea básica es la de comparar los datos observados con la hipótesis que se pone a prueba. Fisher ideó una manera de medir el grado de incompatibilidad de un conjunto de datos con la hipótesis nula, evaluando la probabilidad de hallar resultados como los observados o más extremos, si la hipótesis nula fuera cierta. Si esa probabilidad es muy pequeña, puede suceder que la muestra que se seleccionó haya sido excepcional, o bien que la afirmación hipotética sea falsa. Fisher argumentó que se trataba de un método objetivo para poner a prueba teorías y que puede ser usado en diferentes campos de conocimiento.

Con posterioridad a Fisher, Jerzy Neyman (1894-1981) e Eagon Pearson (1895-1980) introducen dos cambios importantes en el procedimiento.

El primero consiste en tratar a las prueba no ya como métodos para validar teorías, sino como reglas de decisión, es decir, criterios que permiten decidir en las situaciones en que no se cuenta con toda la información necesaria.

El segundo cambio consiste en oponer a la hipótesis nula, otra hipótesis, llamada hipótesis alternativa, a la que se indica como H_1, que es hacia la que se suma evidencia cuando se rechaza H_0.

Veamos la aplicación de este modelo, que es el que usaremos a partir de ahora. La afirmación "los niños cuyos padres tienen alto nivel de educación tienen rendimiento en la escuela superior al promedio general" es una hipótesis, porque pretende tener carácter general, hace referencia a la población de niños en la escuela, los que asisten ahora y los que asistirán en el futuro; no podemos observar a la población completa, por lo tanto la hipótesis no puede probarse de manera definitiva, solo puede hacerse a partir de una muestra. Para formalizar esa hipótesis, construiremos una H_0 que niegue cualquier diferencia: "el rendimiento de niños con padres de alto nivel de educación es el mismo que el del promedio". A esta hipótesis,

opondremos otra, que afirme "los niños cuyos padres tienen alto nivel de educación tienen rendimiento superior al promedio". A esta última llamaremos hipótesis alternativa, H_1. Así formalizamos el planteo del problema.
Supongamos ahora que conocemos ese rendimiento promedio, medido por el puntaje en las pruebas y que vale 60 puntos para la población completa de alumnos. De modo que podemos formular las hipótesis ahora así:

H_0: "El rendimiento promedio de los niños con padres de alto nivel de educación es de 60 puntos"
H_1: "El rendimiento promedio de los niños con padres de alto nivel de educación es superior a 60 puntos"

Para poner a prueba la hipótesis tomaremos una muestra de niños con padres de mucha educación y veremos si su rendimiento es superior a 60 puntos. Supongamos que en la muestra hallamos una media de 62 puntos, ¿estamos autorizados para rechazar la H_0? Aunque 62 es mayor que 60, una diferencia de solo 2 puntos parece demasiado pequeña y podríamos atribuirla al azar. En el razonamiento de la prueba de hipótesis nos interesa evaluar la probabilidad de ocurrencia del resultado que se observa, si la hipótesis nula fuera cierta. En este caso será: "Si los hijos de padres con alto nivel de educación tuvieran el mismo rendimiento que el promedio (H_0), una diferencia de 2 puntos es probable, es esperable, puede deberse a la variabilidad propia de los datos muestrales, en consecuencia, esa diferencia no es suficiente para rechazar la H_0". En otros términos "Si los hijos de padres con alto nivel de educación tuvieran un rendimiento promedio de 60 puntos, no es improbable que una muestra arroje un resultado de 62 puntos". Dicho de otro modo: el resultado muestral no se aleja tanto de lo que esperaríamos si la H_0 fuera cierta, por lo tanto, no podemos rechazarla y concluimos que los hijos de padres con mucha educación no difieren del promedio.
Un elemento de mucha importancia es recordar que la hipótesis hace referencia a la población, mientras que nuestra observación es muestral, y sabemos que los resultados muestrales difieren de los valores paramétricos porque son variables aleatorias.
¿Qué habría sucedido si hubiésemos observado que el grupo de los hijos de padres con mucha educación tienen un rendimiento de 95 puntos? Este resultado se aleja mucho de 60 que es el que sostiene H_0, es decir, si el hipotético fuera verdadero, sería muy

poco probable hallar una muestra que promedie 95 puntos. En consecuencia seguramente rechazaríamos la H_0.
Al realizar pruebas de hipótesis, en lugar de evaluar intuitivamente si un valor muestral está cerca o lejos del valor hipotético, lo que haremos será evaluar cuál sería la probabilidad de hallarlo si fuera cierta la hipótesis nula. Cuando esta probabilidad sea grande no habrá evidencia para rechazarla, cuando sea pequeña decidiremos rechazarla. ¿Cuán grande o pequeña? Es de lo que nos ocuparemos a continuación.

Para realizar una prueba de hipótesis, necesitamos calcular la probabilidad del valor observado, si H_0 fuera cierta, es decir, si el parámetro tuviera ese valor (el que señala H_0). Es una probabilidad condicional que podemos por ahora escribir así:

$$P(^{observado}/_{la\ hipótesis\ nula\ es\ verdadera})$$

Luego haremos más precisa esta expresión.

Comenzaremos con una prueba de hipótesis sobre la media de una variable cuantitativa y luego acerca de la proporción para una categoría de una variable nominal, y lo haremos a través de ejemplos.

Prueba sobre la media

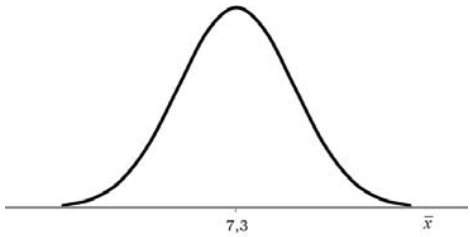
Ejemplo 10.1
Para una determinada carrera universitaria, históricamente los alumnos han tardado para recibirse un promedio de 7,30 años. Decimos históricamente para indicar que son datos acumulados por largo tiempo y que provienen de los registros de la facultad de años atrás. Se ha introducido un cambio en el plan de estudios de la carrera y puede creerse que con ese cambio los alumnos tardarán un tiempo distinto en recibirse. Tenemos entonces un promedio de la población de quienes se recibieron en las anteriores condiciones (una media poblacional histórica), y queremos hacer inferencia sobre la media poblacional de los alumnos que cursan con el nuevo plan. Estos últimos no están todos accesibles, porque hay alumnos que están cursando y otros que lo harán en el futuro, por lo que de esa población solo puedo conocer a una muestra de los que ya han egresado y ver cuánto tiempo han tardado ellos en recibirse.
Expresamos las hipótesis de este modo:

$$H_0: \mu = 7{,}30$$

$$H_0: \mu \neq 7{,}30$$

La hipótesis nula indica que la media poblacional de los alumnos que cursan con el nuevo plan es la misma que antes, que no hay diferencia, que no hay cambios. La hipótesis alternativa afirma lo contrario: que el tiempo promedio que tardan los alumnos en terminar la carrera con el nuevo plan es diferente a los 7,30 años históricos. Ambas son afirmaciones sobre la población (sobre el parámetro media poblacional), por eso son hipótesis.
Si la hipótesis nula fuera cierta, por lo que sabemos sobre las distribuciones en el muestreo, la siguiente sería la distribución de las medias muestrales:

Gráfico 1: Distribución de las medias muestrales bajo la hipótesis nula



Que quiere decir que "lo más probable" sería hallar a la media muestral alrededor de 7,30. Hay poca probabilidad de encontrar valores muy lejanos a 7,30, como lo muestran las áreas decrecientes, a medida que nos alejamos de la media hipotética. Por esta razón, para decidir si un resultado muestral se aleja mucho o poco del valor paramétrico, deberemos determinar si es poco probable o muy probable. Será equivalente decir que un valor se aleja mucho de la media hipotética que decir que se trataría de un valor poco probable, si la media fuera la que propone la H_0.

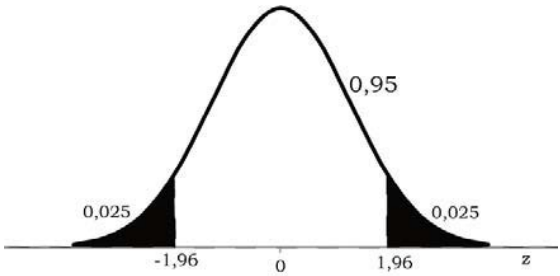
Afirmar de un valor muestral que **se aleja mucho del valor poblacional** equivale a decir que **sería muy poco probable si el valor poblacional fuera el hipotético**

A fin de realizar la prueba de hipótesis debemos obtener una muestra. Supongamos que seleccionamos 100 egresados (usando un muestreo irrestricto aleatorio) y que encontramos un tiempo promedio para terminar la carrera de 7,50 años con una

desviación estándar de 1,30 años ($\bar{x} = 7{,}50\ y\ s = 1{,}30$). Debemos tener un criterio para decidir si este valor observado es compatible con la hipótesis nula ($\mu = 7{,}30$) o si constituye evidencia suficiente para rechazarla a favor de la hipótesis alternativa ($\mu \neq 7{,}30$). El criterio es el de ver cuán probable sería este valor observado si la hipótesis nula fuera cierta. En consecuencia, debemos calcular la probabilidad que tiene $\bar{x}$ de asumir el valor observado. Sin embargo, no es posible hallar probabilidades para valores únicos de una variable continua, por lo que buscaremos la probabilidad de hallar valores como el observado (7,50) *o más extremos que él*. Esto significa que nos preguntamos por la probabilidad que tiene la variable $\bar{x}$ de asumir el valor 7,50 *o uno más extremo*, es decir un valor que se aleje más de la media hipotética. Hemos dicho que "alejado" equivale a "poco probable si H_0 fuera cierta", por lo que los valores alejados se encuentran en los extremos de la distribución de $\bar{x}$, bajo el supuesto de H_0 verdadera (es decir, centrada en la media hipotética).
Para decidir si la evidencia hallada en la muestra es suficiente para rechazar la hipótesis nula, vamos a establecer a priori un valor máximo para la probabilidad de ocurrencia del valor muestral, o lo que es lo mismo, un valor máximo para el área extrema donde consideraremos que se encuentran los valores "alejados".

La toma de decisión
La H_0 se rechazará si hay poca probabilidad de hallar un valor como el observado o uno más extremo que él. Lo que llamamos poca probabilidad, puede establecerse a priori, por ejemplo en 0,05. Eso indica que consideraremos a los resultados con probabilidad menor a 0,05 como muy improbables de hallar si H_0 fuera cierta y nos conducirán a rechazarla. Por el contrario, si encontramos valores cuya probabilidad de ocurrencia es superior a 0,05, los trataremos como valores esperables y nos conducirán a aceptar la H_0. Como sabemos de la distribución normal, los valores de $z = \pm 1{,}96$ delimitan un área central de 95%, es decir que dejan fuera un área de 5%. Los valores de z superiores a 1,96 ó inferiores a -1,96 tienen una probabilidad de ocurrencia de 0,05, repartida en las dos "colas" de la distribución normal.

Gráfico 2: Áreas extremas que totalizan una probabilidad de 0,05



Los valores de $\bar{x}$ que correspondan a puntajes z que superen a 1,96 ó sean inferiores a -1,96 serán valores con probabilidad menor a 0,05, por lo que serán considerados como poco probables y conducirán a rechazar H_0. Por el contrario, los valores que tengan z comprendido entre -1,96 y 1,96 serán probables y nos llevarán a que aceptemos H_0. Estos dos puntos (-1,96 y 1,96) se denominan **valores críticos** de z y se indican con un subíndice: z_c.
En nuestro ejemplo, el valor observado es $\bar{x} = 7{,}50$, de aquí en adelante lo llamaremos $\bar{x}_{obs}$. El puntaje z equivalente a ese $\bar{x}_{obs}$ se llama z observado (z_{obs}) y vale:

$$z_{obs} = \frac{\bar{x}_{obs} - \mu}{\frac{s}{\sqrt{n}}} = \frac{7{,}50 - 7{,}30}{\frac{1{,}30}{\sqrt{100}}} = 1{,}54$$

Se trata de la transformación a puntaje z del valor observado de la media muestral. Se conoce con el nombre de **estadístico de prueba**.
Este puntaje no está en la zona extrema, porque no va más allá de 1,96; por el contrario, está entre -1,96 y 1,96 que pertenece a la parte de valores centrales de la distribución, los más probables. En consecuencia, la decisión es la de aceptar Ho y concluir que el tiempo que los alumnos tardan en completar la carrera no ha cambiado respecto del valor histórico. Dicho de otra manera, el valor observado de $\bar{x}_{obs} = 7{,}50$ es un resultado esperable si la media poblacional fuera de 7,30.
Por la forma en que hemos razonado y tomado la decisión, se comprende que a los valores de z comprendidos entre -1,96 y 1,96 se los denomine **zona de aceptación de H_0**. El otro conjunto de valores de z, los mayores a 1,96 junto a los menores a -1,96, constituyen la **zona de rechazo de H_0**. Luego de haber considerado a 0,05 como la probabilidad a la que llamamos

"pequeña", quedaron determinados los valores de z_c que indican las zonas de aceptación y de rechazo.

La **zona de rechazo de H_0** es el conjunto de valores extremos de la distribución, donde es poco probable encontrar los valores muestrales si H_0 es verdadera.

La **zona de aceptación de H_0** es el conjunto de valores centrales de la distribución, donde es más probable encontrar los valores muestrales si H_0 es verdadera.

Luego de eso, el procedimiento que seguimos fue: calcular el puntaje z que corresponde al valor observado de $\bar{x}$, y luego ver si éste se encuentra en la zona de aceptación o de rechazo de H_0.

La probabilidad 0,05 como valor pequeño fue una elección y podría haber sido diferente; ese número tiene una larga tradición histórica, Fisher lo usaba regularmente, aunque aclarando que no era obligatorio y que no hay nada especial para elegirlo[70]. Se conoce como **nivel de significación** y se indica con la letra α. Es la probabilidad de hallar un valor como el observado o más extremo que él, si la hipótesis nula fuera cierta, por lo que es una probabilidad condicional que ahora escribimos como:

$$P(z<-1,96 \text{ ó } z>1,96/H_0 \text{ es verdadera})=0,05$$

De esta expresión es importante recordar que alfa mide la probabilidad de hallar a z en la región de rechazo (más allá de los puntos críticos) si H_0 es verdadera.

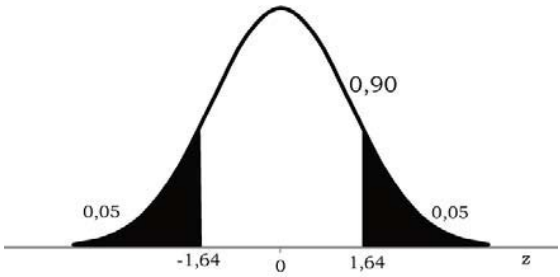
El valor que elijamos para alfa indica a qué valores vamos a considerar como poco probables: en este caso se trata de valores tan poco probables como el 5%. Puede usarse un nivel de significación diferente, por ejemplo del 10% y los valores críticos de z serán diferentes. En efecto los puntos que dejan un área extrema del 10% son $z_c=\pm 1,64$.
En ese caso escribiremos:

$$P(z<-1,64 \text{ ó } z>1,64/H_0 \text{ es verdadera})=0,10$$

[70] En 1965, en *Statistical methods and scientific inference*, Fisher señaló: "Ningún investigador tiene un nivel de significación fijo, al cual año tras año y en toda circunstancia rechaza hipótesis; más bien entrega su mente a cada caso particular a la luz de la evidencia y de sus ideas"

Se llama **nivel de significación** a la probabilidad de hallar al valor muestral en la zona de rechazo de H_0, si H_0 es verdadera. Se indica como **α**, y es elegido por el investigador

Gráfico 3: Áreas extremas que totalizan una probabilidad de 0,10



Por lo que, si el nivel de significación es 0,10 (α=0,10), la zona de aceptación de H_0 es el conjunto de valores *z* comprendidos entre -1,64 y 1,64 (centrales), mientras que la zona de rechazo de H_0 son los *z* menores a -1,64 y los mayores a 1,64 (los valores extremos cuya probabilidad es el área sombreada en el gráfico). El valor muestral del ejemplo ($z_{obs} = -1{,}54$) está también en la zona de aceptación para este nivel, por lo que tampoco se rechaza la H_0 a un nivel de significación de 0,10.

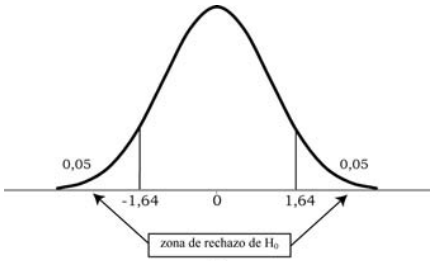
Otro nivel de significación que suele usarse es del 1%. Para él, los valores de *z* son ±2,56, por lo que la regla de decisión será: "si el valor de *z* correspondiente al valor observado de $\bar{x}$ está entre -2,56 y 2,56 se debe aceptar la H_0, si es menor a -2,56 ó superior a 2,56 se debe rechazar la H_0".
Expresamos la probabilidad condicional como:

$$P(z<-2{,}56 \text{ ó } z>2{,}56/H_0 \text{ es verdadera})=0{,}01$$

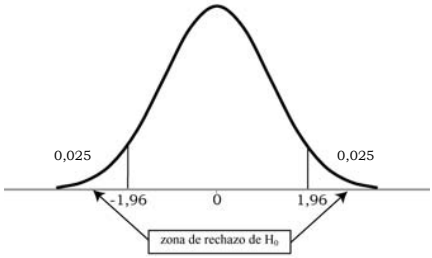
Al igual que en los casos anteriores, cuando se expresan en términos de puntajes *z*, estos valores son fijos; no dependen de los resultados muestrales que se encuentren, constituyen una regla de decisión establecida a priori. Cuanto más pequeño se elije α, tanto más exigente es la prueba, en el sentido de que solo rechaza la hipótesis de no-diferencia si se observan valores muy alejados del hipotético.

Gráfico 4: Comparación de las zonas de aceptación y rechazo de H_0 con niveles de significación del 10, del 5 y del 1%

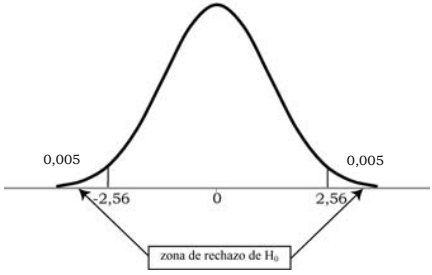
$\alpha = 0{,}10$



$\alpha = 0{,}05$



$\alpha = 0{,}01$



Ejemplo 10.2
Para la misma carrera universitaria del ejemplo anterior, el promedio de calificación con que terminaban los estudios los alumnos era, según registros históricos, de 6,50. Nos preguntamos si, luego del cambio en el plan de estudios, esta nota promedio ha cambiado o sigue siendo la misma. El planteo de las hipótesis será ahora:

$$H_0: \mu = 6{,}50$$
$$H_0: \mu \neq 6{,}50$$

Por la misma razón que antes, no podemos analizar a la población completa, usaremos los datos obtenidos en una muestra. En ella encontramos, por ejemplo, que el promedio de

los 100 egresados es de 6,65 con desviación estándar de 0,60, es decir: $\bar{x} = 6{,}65\ y\ s = 0{,}60$. A un nivel de significación del 5%, los puntos críticos vuelven a ser z_c = ±1,96. Buscamos el estadístico de prueba transformando el valor observado de $\bar{x}$ a puntaje z y encontramos:

$$z_{obs} = \frac{\bar{x}_{obs} - \mu}{\frac{s}{\sqrt{n}}} = \frac{6{,}65 - 6{,}50}{\frac{0{,}6}{\sqrt{100}}} = 2{,}50$$

El valor observado de $\bar{x}$ corresponde entonces a un z que supera al punto crítico (que es z_c=1,96), por lo que está en la zona de rechazo de H_0. La decisión es rechazar H_0 y concluir que el promedio de los alumnos es actualmente diferente del promedio histórico.

En los dos ejemplos vemos que la regla de decisión depende del nivel de significación. Cuando se fija en el 5% entonces se puede expresar como "si el valor de z correspondiente al valor observado de $\bar{x}$ está entre -1,96 y 1,96 se debe aceptar la H_0, si es menor a -1,96 ó superior a 1,96 se debe rechazar la H_0". Cuando el nivel de significación es del 10%, diremos que "si el valor z correspondiente al valor observado de $\bar{x}$ está entre -1,64 y 1,64 se debe aceptar la H_0, si es menor a -1,64 ó superior a 1,64 se debe rechazar la H_0".

Veamos más en detalle el significado de esta probabilidad que hemos fijado en 0,05 y que puede también elegirse en 0,10 ó en 0,01 y que llamamos α. Se trata de la probabilidad de hallar el valor observado en la muestra (o uno más extremo a él) si la H_0 fuera verdadera, por lo que cada vez que hallemos valores muestrales que se encuentran allí, tomaremos la decisión de rechazar H_0. Si la hipótesis nula fuera efectivamente verdadera, la decisión sería incorrecta, pero a eso no lo sabemos, porque nunca conocemos el verdadero valor del parámetro. Aunque sí podemos afirmar que al fijar α en el 5%, ésas serán las chances de equivocarnos rechazando una hipótesis nula que era verdadera. En el segundo ejemplo, cuyo resultado fue el de rechazar H_0, es muy importante indicar a qué nivel de significación se toma la decisión, porque ese número (5%) indica la probabilidad de haber tomado la decisión erróneamente. Mide la probabilidad de haber encontrado el promedio muestral de 6,65 por azar. Como esa probabilidad es pequeña, decidimos rechazar H_0.

Los puntos críticos en términos del estimador
Hay una manera diferente de establecer las zonas de aceptación y rechazo, que consiste en fijar los puntos críticos en términos de $\bar{x}$, en lugar de hacerlo como puntajes *z*. Por lo que en lugar de determinar los dos z_c, hallaremos los dos valores críticos de $\bar{x}$, a los que llamaremos $\bar{x}_c$[71]:

$$\bar{x}_c = \mu \pm z_c * \frac{s}{\sqrt{n}}$$

En el ejemplo 11.1 (sobre el tiempo que tardan los alumnos en terminar la carrera) y a un nivel de significación de 5%, los valores de $\overline{x_c}$ son:

$$\bar{x}_c = \mu \pm z_c * \frac{s}{\sqrt{n}} = 7{,}30 \pm 1{,}96 * \frac{1{,}30}{\sqrt{100}} = 7{,}30 \pm 0{,}25$$

Al sumar obtenemos 7,55 y al restar 7,05. Estos son los puntos críticos expresados en términos de la variable original. La regla de decisión es ahora "si se encuentra un valor de $\bar{x}_{obs}$ comprendido entre 7,05 y 7,55 se debe aceptar la H_0. Si el valor observado de $\bar{x}_{obs}$ es inferior a 7,05 ó superior a 7,55 se debe rechazar H_0.
Para expresarlo como probabilidad condicionada:

$$P(\bar{x}<7{,}05 \text{ ó } \bar{x}>7{,}55)/\mu=7{,}30)=0{,}05$$

Que afirma que la probabilidad de hallar a $\bar{x}$ por debajo de 7,05 ó por encima de 7,55 si la media de la población es 7,30, vale 0,05.

Al hacer la prueba, vemos que $\bar{x}_{obs} = 7{,}50$, que no va más allá de los puntos críticos, por lo que pertenece a la zona de aceptación de H_0. Concluimos que se acepta H_0 y que los alumnos no han cambiado el tiempo que tardan en terminar la carrera. La regla de decisión es la misma que antes, solo que ahora está expresada en el lenguaje de $\bar{x}$ y no de *z* y la conclusión también es la misma.

Volviendo ahora sobre el caso de los promedios con que egresan los estudiantes (ejemplo 11.2), para hallar los valores críticos de la media muestral hacemos:

[71] Aunque esta expresión es parecida a la de los intervalos de confianza, no se deben confundir. Aquí el centro está en el valor hipotético del parámetro, mientras que los intervalos de confianza se centran en el valor del estimador puntual.

$$\bar{x}_c = \mu \pm z_c * \frac{s}{\sqrt{n}} = 6{,}50 \pm 1{,}96 * \frac{0{,}60}{\sqrt{100}} = 6{,}50 \pm 0{,}12$$

Y resultan: 6,38 y 6,62. El promedio observado fue de 6,65, que supera al punto crítico superior y se encuentra en la zona de rechazo. Concluimos que se rechaza la H_0 y los alumnos egresan en la actualidad, con un promedio que difiere del histórico. Nuevamente, es la misma conclusión que si se trabaja sobre *z*.

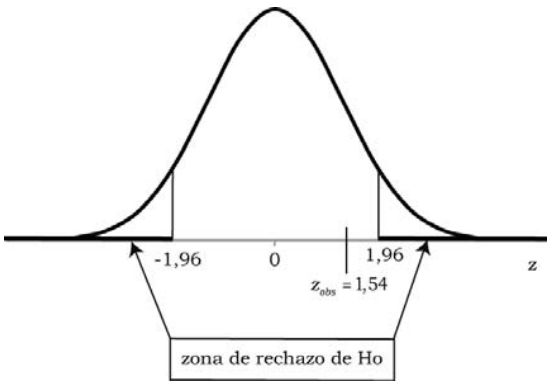
Comparemos los dos procedimientos:
A. Usando valores críticos de z

1. Habiendo establecido el nivel de significación, determinar los valores z que dejan esa probabilidad en los extremos. Éstos son los z_c.
2. Para hallar el estadístico de prueba, transformar el valor observado de $\bar{x}$ en puntaje z haciendo:

$$z_{obs} = \frac{\bar{x}_{obs} - \mu}{\frac{s}{\sqrt{n}}}$$

3. Observar la posición de este valor transformado en la distribución de probabilidades z
Por ejemplo, para un nivel de significación de 0,05 (ó 5%), en el ejemplo 11.1 resulta:

Gráfico 5: Ubicación de la zona de rechazo de H_0 a un nivel de significación de 0,05, sobre puntajes estándar (*z*), y del valor observado.



B. Usando los valores críticos de $\bar{x}$
1. Habiendo establecido el nivel de significación, determinar los valores z que dejan esa probabilidad en los extremos. Éstos son los z_c.

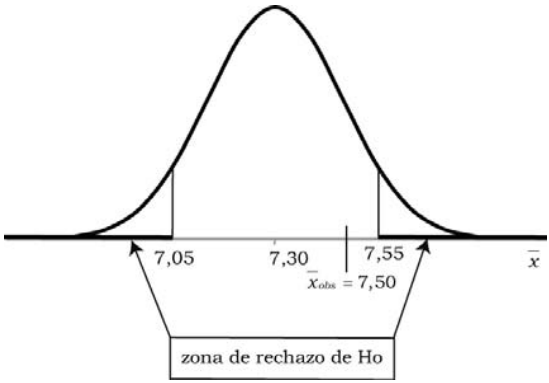
2. Usar los z_c para determinar los correspondientes $\bar{x}_c$ haciendo:

$$\bar{x}_c = \mu \pm z_c * \frac{s}{\sqrt{n}}$$

3. Ver la posición de $\bar{x}_{obs}$ en la distribución de probabilidades de $\bar{x}$.

Por ejemplo, para el nivel de significación de 0,05 y el mismo ejemplo, resulta:

Gráfico 6: Ubicación de la zona de rechazo de H_0 a un nivel de significación de 0,05, sobre valores de la variable ($\bar{x}$), y del valor observado



La diferencia entre las dos formas de establecer los puntos críticos es que con la primera se determinan los valores de z_c a partir del nivel de significación y luego se transforma a puntaje z el valor muestral observado de $\bar{x}_{obs}$. En el segundo modo, los z_c se transforman (al revés) en puntos críticos de $\bar{x}_c$ y luego se compara el $\bar{x}_{obs}$ directamente, sin transformarlo.

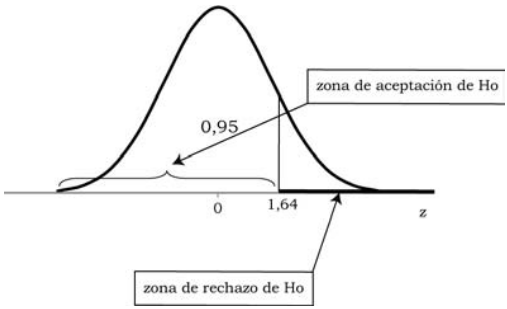
Los gráficos 5 y 6 expresan lo mismo, el primero en el lenguaje de estandarizado de z, el segundo en el de $\bar{x}$. Los procedimientos son equivalentes y puede vérselos aplicados de manera indiferenciada.

Pruebas unilaterales

A menudo la hipótesis alternativa no expresa solo que la media difiera del valor hipotético, sino que indica en qué *dirección* se espera que difiera. Por ejemplo, puede esperarse, en el ejemplo anterior, que los alumnos egresen con promedio superior a 6,50, con lo que ahora la H_1 dirá que $\mu > 6{,}50$. Se trata en este caso de una prueba unilateral y solo rechazaremos la H_0 si encontramos valores sustancialmente *mayores* que 6,50. Para el mismo nivel de significación del 5%, el valor *z* que nos interesa es el que delimita un área *superior* de 0,05.

Notemos la diferencia con las pruebas que tratamos antes: al nivel de 5% buscábamos dos *z* que dejaban en total 0,05 de área extrema (por encima y por debajo) o lo que es lo mismo, los dos *z* que dejan el 95% del área central. Ahora, como la prueba es unilateral y solo nos interesan valores que se excedan, solo buscamos un *z*, el que deja al 5% por encima. Ese valor de *z* es 1,64.[72]

Gráfico 7: Ubicación de las zonas de aceptación y rechazo de Ho para una prueba unilateral derecha.



Por oposición a las anteriores, las pruebas unilaterales se llaman **pruebas de una cola**. Como vemos en el gráfico, el conjunto de valores *z* que conducen a rechazar H_0 se encuentran solo a la derecha.

Ejemplo 10.3

En la situación en que nuestro interés esté en analizar si el promedio con que egresan ahora los estudiantes es *superior* al valor histórico (y no solo diferente a él), las hipótesis de la prueba se expresan:

[72] Este número es el mismo que usamos al nivel del 10% en pruebas bilaterales. Esto se debe a que en ese caso el 10% extremo se reparte en 5% en cada cola; ahora nos interesa una sola cola, del 5%.

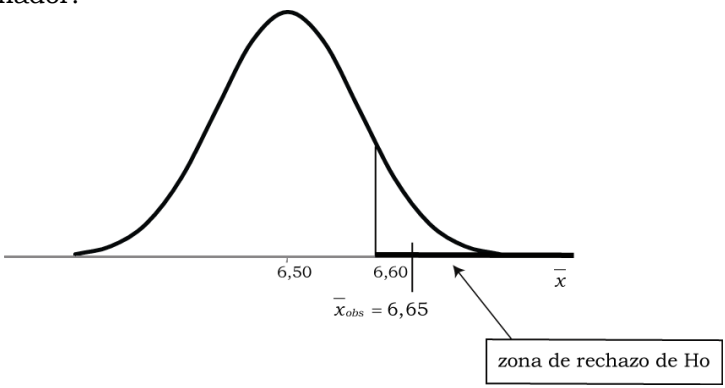
$$H_0: \mu = 6{,}50$$
$$H_1: \mu > 6{,}50$$

Haremos la prueba sobre los valores de $\bar{x}$, a un nivel de significación del 5% entonces:

$$\bar{x}_c = \mu + z_c * \frac{s}{\sqrt{n}} = 6{,}50 + 1{,}64 * \frac{0{,}60}{\sqrt{100}} = 6{,}50 + 0{,}10$$

Al sumar obtenemos 6,60 que es el único punto crítico que nos interesa por tratarse de una prueba unilateral y hemos sumado porque la prueba es derecha. Con lo que resulta la siguiente región de rechazo de H_0:

Gráfico 8: Región de rechazo unilateral derecha de $H_0: \mu = 6{,}50$ a un nivel de significación del 5%, con una muestra de 100 casos y desviación estándar 0,60, y ubicación del valor observado del estimador.



El valor observado de $\bar{x}$ había sido 6,65 que es mayor que el punto crítico, con lo que rechazamos la H_0 y concluimos que los alumnos egresan con un promedio significativamente mayor al histórico.

Si hubiésemos planteado la regla de decisión sobre los valores de z, a un nivel de significación del 5%, corresponde observar el gráfico 7 y obtener el puntaje z correspondiente al valor observado de la media muestral:

$$z_{obs} = \frac{\bar{x}_{obs} - \mu}{\frac{s}{\sqrt{n}}} = \frac{6{,}65 - 6{,}50}{\frac{0{,}60}{\sqrt{100}}} = 2{,}5$$

Dado que este resultado supera al punto crítico del gráfico 7, concluimos nuevamente que los alumnos egresan con un promedio significativamente mayor al histórico.

En los párrafos anteriores hemos introducido una expresión nueva, que tiene un sentido preciso. Cuando decimos "significativamente mayor" no nos referimos al uso que suele darse en el lenguaje cotidiano, que es sinónimo de importante, de gran magnitud, grande, etc. Diremos que un valor es significativamente mayor o menor si se ha rechazado una prueba unilateral, o bien que es significativamente diferente o que la diferencia es significativa, si la H_0 que fue rechazada pertenece a una prueba bilateral. Además debe indicarse el nivel de significación de la prueba, por lo que diremos: según los datos observados y a un nivel del 5%, los alumnos egresan con un promedio significativamente superior al histórico.

Un resultado es **significativo** cuando conduce a rechazar una H_0 a un determinado nivel de significación.

Un resultado puede ser significativo a un nivel y no serlo a otro. Por ejemplo, si en una prueba bilateral y luego de transformar el valor observado a puntaje z, se obtiene z=2,3, este resultado conducirá a que se rechace Ho al 5% (porque 2,3 es mayor que 1,96) pero que no se rechace al 1% (porque 2,3 no es mayor que 2,56). En ese caso diremos que se obtienen resultados significativos al 5% pero no al 1%. Luego veremos que esta clasificación puede hacerse más precisa.

Otros ejemplos de prueba de hipótesis sobre la media:
La investigación con sujetos animales en Psicología se basa en que distintas especies comparten mecanismos básicos, y los animales no humanos presentan un nivel de complejidad menor, lo que facilita la comprensión de complejos fenómenos en humanos. De hecho, los *modelos animales* han permitido avanzar en el conocimiento de mecanismos neurofisiológicos, cognitivos y comportamentales de seres humanos, siendo clave en el ámbito de la psicopatología y terapia psicológica.
En este contexto, alumnas de Psicología interesadas en la vulnerabilidad adolescente hacia el uso y abuso de drogas, recurrieron a un modelo animal para estudiar los efectos del alcohol[73]. Uno de los posibles factores que determinan las primeras aproximaciones a las drogas es la búsqueda de nuevas sensaciones, que se manifiesta en niveles elevados tanto en adolescentes humanos como no humanos, particularmente en roedores. En modelos animales, una manera de indagar el comportamiento de búsqueda de novedad es a través del uso de

[73] Dziula y Reyna (2005)

un objeto novedoso. Debido al escaso conocimiento sobre el comportamiento de los roedores ante tales objetos, las alumnas desarrollaron un estudio piloto. Concretamente, expusieron a 32 ratas Wistar[74] adolescentes (28 a 42 días de edad) a un objeto novedoso durante 3 minutos en distintos intervalos de tiempo.
A continuación, se retoman datos parciales del estudio piloto[75] a los fines de exponer la modalidad de trabajo cuando se realiza una prueba de hipótesis sobre un valor determinado de la media. Primero ilustramos el caso de una prueba bilateral y luego unilateral.

Ejemplo 10.4
En algunos estudios se ha observado que el tiempo que tardan los roedores (en general) en contactar un objeto novedoso (llamado tiempo de latencia, o simplemente latencia) es de 150 seg. En función de ello, se indaga si la latencia de contacto con el objeto en ratas Wistar adolescentes es la misma o no en relación a la reportada previamente en otros trabajos con roedores. Las hipótesis formuladas son:

$$H_0: \mu = 150$$
$$H_1: \mu \neq 150$$

En la muestra utilizada, la latencia media de contacto con el objeto es de 138,42 seg. y la desviación estándar 42,98. Se establece un nivel de significación del 5%, por lo que los puntos críticos en puntajes z son $z_c = \pm 1{,}96$.
Luego, transformamos el valor de $\bar{x}$ observado en puntaje z a través de la siguiente fórmula:

$$z_{obs} = \frac{\bar{x}_{obs} - \mu}{\frac{s}{\sqrt{n}}} = \frac{138{,}42 - 150}{\frac{42{,}98}{\sqrt{32}}} = \frac{-11{,}58}{7{,}59} = -1{,}52$$

Vemos que al valor de $\bar{x}$ observado le corresponde un valor z que se encuentra entre los puntos críticos $z \pm 1{,}96$, es decir que se halla en la zona de no rechazo de la H_0, por lo que concluimos que no hay evidencia que indique que la latencia media de contacto con un

[74] Se trata de una línea albina de la rata parda. Fue desarrollada en el Wistar Institute en 1906 para fines de investigación biomédica, y se trata de la primera rata empleada como organismo modelo (anteriormente se trabajaba con el ratón).
[75] Algunos de los datos son ficticios debido a la falta de disponibilidad y a la necesidad de adaptar el ejemplo a nuestros fines. Se mantiene la temática del estudio de referencia, aunque los resultados mostrados no corresponden exactamente a él.

objeto novedoso en ratas Wistar es distinta a la latencia que manifiestan los roedores en general.

Se obtiene la misma conclusión si se utilizan los valores críticos de $\bar{x}$. Revisemos el procedimiento: luego de haber establecido el nivel de significación al 5%, se obtienen los puntos críticos en puntajes originales ($\bar{x}_{obs}$) a partir de los puntos críticos en puntajes z (z_c), a través de la siguiente fórmula[76]:

$$\bar{x}_c = \mu \pm z_c * \frac{s}{\sqrt{n}} = 150 \pm 1{,}96 * \frac{42{,}98}{\sqrt{32}} = 150 \pm 14{,}89$$

Entonces, los $\bar{x}_c$ que delimitan las zonas de rechazo y no rechazo son 135,02 y 164,89, y el valor observado de $\bar{x} = 138{,}42$ se encuentra comprendido entre ellos, por lo que no se rechaza la H_0.

Ejemplo 10.5
La literatura sobre comportamiento exploratorio en roedores (en general) indica que cuando los organismos son expuestos durante 180 seg a un objeto novedoso permanecen en contacto con el mismo (duración) 14 seg en promedio. Debido a las características del período adolescente, en este trabajo se formula una hipótesis que indica que las ratas adolescentes estarán más tiempo en contacto con el objeto. Las hipótesis formuladas son:

$$H_0: \mu = 14$$
$$H_1: \mu > 14$$

En los animales que comprenden la muestra bajo análisis, se observa que la duración de contacto con el objeto es de 17,43 seg. y la desviación estándar 4,25. El nivel de significación se establece en el 5% y, dado que la prueba es unilateral derecha, el punto crítico en puntaje z es 1,64.
Luego, transformamos el valor de $\bar{x}$ observado en puntaje z a través de la siguiente fórmula:

$$z_{obs} = \frac{\bar{x}_{obs} - \mu}{\frac{s}{\sqrt{n}}} = \frac{17{,}43 - 14}{\frac{4{,}25}{\sqrt{32}}} = \frac{3{,}43}{0{,}75} = 4{,}57$$

Al valor de $\bar{x}$ observado le corresponde un valor z = *4,57*, que resulta superior al z_c = *1,64*, por lo que se rechaza la H_0, es decir

[76] Sumamos y restamos para obtener dos puntos críticos, porque se trata de una prueba bilateral.

que las ratas adolescentes muestran una duración mayor de contacto con un objeto novedoso que lo señalado por la literatura para roedores en general.
Obtenemos la misma conclusión si usamos los valores críticos de $\bar{x}$. Una vez establecido el nivel de significación al 5%, se obtienen los puntos críticos en puntajes originales ($\bar{x}_c$) a partir de los puntos críticos en puntajes z (z_c), a través de la siguiente fórmula[77]:

$$\bar{x}_c = \mu + z_c * \frac{s}{\sqrt{n}} = 14 + 1{,}64 * \frac{4{,}25}{\sqrt{32}} = 14 + 1{,}23 = 15{,}23$$

El valor observado de $\bar{x} = 17{,}43$ resulta mayor a $\bar{x}_c = 15{,}23$, por lo que se rechaza la H_0, y concluimos que la duración promedio de contacto con un objeto novedoso en ratas adolescentes es significativamente mayor a 14.

Ejemplo 10.6
Luego de haber evaluado la latencia de contacto con un objeto novedoso de los animales en repetidas ocasiones, los investigadores observan un valor promedio de 112,48 seg. Ahora, están interesados en indagar cuál será la latencia de contacto si las ratas son nuevamente expuestas al objeto novedoso, suponiendo que será menor debido a la disminución del carácter novedoso que hacía que los animales tardaran en contactar el objeto en las exposiciones iniciales. Las hipótesis formuladas son:

$$H_0: \mu = 112{,}48$$
$$H_1: \mu < 112{,}48$$

En la nueva exposición, la latencia promedio de contacto con el objeto es de 79,53 y la desviación estándar es de 41,22. El nivel de significación se establece en el 5% y, dado que la prueba es unilateral izquierda, el punto crítico en puntaje z es -1,64.
Luego, transformamos el valor de $\bar{x}_c$ observado en puntaje z calculando el estadístico de prueba:

$$z_{obs} = \frac{\bar{x}_{obs} - \mu}{\frac{s}{\sqrt{n}}} = \frac{79{,}53 - 112{,}48}{\frac{41{,}22}{\sqrt{32}}} = \frac{-32{,}95}{7{,}29} = -4{,}52$$

Vemos entonces que al valor de $\bar{x}_c$ observado le corresponde un valor z = *-4,52*, que resulta inferior al z_c = *-1,64*, por lo que se

[77] Solo sumamos para obtener el punto crítico de la derecha, porque se trata de una prueba unilateral derecha.

rechaza la H_0, es decir que la latencia de contacto con el objeto novedoso en una nueva exposición es significativamente menor a 112,48 seg.
A la misma conclusión se arriba si se realizan los cálculos con los valores críticos de $\bar{x}_c$, que se obtienen a partir de los puntos críticos en puntajes z (z_c), a través de la siguiente fórmula[78]:

$$\bar{x}_c = \mu - z_c * \frac{s}{\sqrt{n}} = 112{,}48 - 1{,}64 * \frac{41{,}22}{\sqrt{32}} = 112{,}48 - 11{,}95 = 100{,}53$$

El valor observado de $\bar{x} = 79{,}53$ resulta menor a $\bar{x}_c = 100{,}53$, por lo que se rechaza la H_0, la latencia promedio de contacto con el objeto novedoso en la nueva exposición es significativamente menor a 112,48 seg.
Debemos recordar que el carácter unilateral o bilateral de la prueba no depende de la H_0 sino de la H_1. En efecto, la H_0 siempre indica un valor determinado para el parámetro (hasta aquí la media), mientras que la H_1 puede indicar un valor diferente si la prueba es bilateral, o bien señalar la dirección de la diferencia hacia los mayores o menores y en esos casos, la prueba es unilateral. La decisión de hacer una prueba unilateral o bilateral depende de cada investigación concreta, de la pregunta que el investigador formula.

Cuadro 1: Valores críticos usuales de la distribución normal

Significación	Puntaje z para prueba:		
	Bilateral	Unilateral derecha	Unilateral izquierda
0,10	±1,64	+1,28	-1,28
0,05	±1,96	+1,64	-1,64
0,01	±2,57	+2,33	-2,33

[78] Solo restamos para obtener el punto crítico izquierdo, porque es una prueba unilateral izquierda.

Prueba sobre la proporción

De modo equivalente a los intervalos de confianza, hacer una prueba de hipótesis sobre una proporción, conlleva los mismos pasos que cuando se trata de la media. Se plantean, en primer lugar, las hipótesis nula y alternativa. La hipótesis nula afirma un valor para la proporción poblacional, mientras que la hipótesis alternativa puede, o bien solo indicar que el valor es diferente (prueba bilateral), o bien precisar si la diferencia se espera hacia valores mayores o menores que los indicados por la hipótesis nula (prueba unilateral). Una vez fijado el nivel de significación (α) y la lateralidad de la prueba, quedan determinados los puntos críticos en términos de z, según el cuadro 1.

La principal diferencia a tener en cuenta es el cálculo del error estándar de la proporción que, según vimos en la distribución en el muestreo de las $\hat{p}$ es:

$$\sigma_{\hat{p}} = \sqrt{\frac{P * (1 - P)}{n}}$$

Recordemos que cuando construíamos los intervalos de confianza era necesario aproximar $P * (1 - P)$ a través de $\hat{p} * (1 - \hat{p})$, porque no conocíamos el valor de la proporción poblacional. Ahora, la situación es diferente, porque tenemos una P (poblacional) hipotética, y es esa la que usaremos para el cálculo de $\sigma_{\hat{p}}$. Por lo tanto, la transformación del valor observado en la muestra a puntajes z se hará según:

$$z_{obs} = \frac{\hat{p}_{obs} - P}{\sqrt{\frac{P * (1 - P)}{n}}}$$

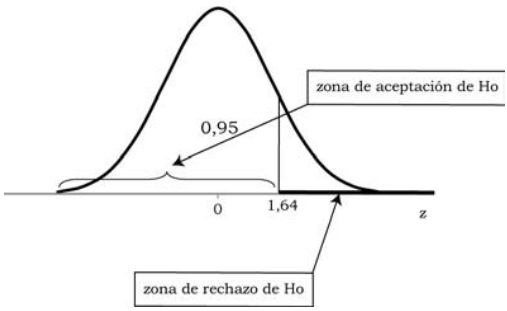
que es el estadístico de prueba para la prueba de proporciones. De acuerdo a que la posición de este z_{obs}, sea en la zona de aceptación o de rechazo de H_0, se toma la decisión.

Ejemplo 10.7

Veamos una aplicación con datos ficticios en el caso de una prueba unilateral derecha. Un político tenía, haces tres meses una intención de voto equivalente al 30% del padrón. De acuerdo con algunas acciones de campaña, creemos que esta proporción pudo haber aumentado, por lo que el planteo de las hipótesis es:

$$H_0: P = 0{,}30$$
$$H_1: P > 0{,}30$$

Al igual que sucedió con la media, estos son valores que planteamos de manera hipotética acerca del parámetro, en este caso la proporción poblacional (*P*). La H_0 indica que la proporción de votos sigue siendo la misma. Se trata de una prueba unilateral derecha, porque estamos interesados en encontrar un eventual aumento en la proporción de votantes que tiene el candidato, por eso la H_1 indica una proporción mayor.
Establecemos un nivel de significación de 5%, por lo que el valor crítico de *z* (en prueba unilateral) es 1,64. En términos de *z*, las zonas de aceptación y rechazo quedan así:



Para poner a prueba la hipótesis analizaremos la intención de voto de una muestra de 200 ciudadanos, en la que hallamos que 65 dicen que votaría a ese candidato. En la muestra entonces:

$$\hat{p} = \frac{65}{200} = 0{,}325$$

Nos preguntamos si este valor puede considerarse como un verdadero aumento respecto del 30% anterior o si solo se explica por razones de azar. Repitiendo la operación que realizamos para la media, transformamos este valor observado de la proporción muestral a puntaje *z* y hallamos el estadístico de prueba:

$$z_{obs} = \frac{\hat{p}_{obs} - P}{\sqrt{\frac{P*(1-P)}{n}}} = \frac{0{,}325 - 0{,}30}{\sqrt{\frac{0{,}30*(1-0{,}30)}{200}}} = 0{,}77$$

Este puntaje de z_{obs} no supera al punto crítico ($z=1{,}64$) por lo que se sitúa en la zona de aceptación de H_0. Concluimos que la proporción no ha aumentado respecto del valor anterior.

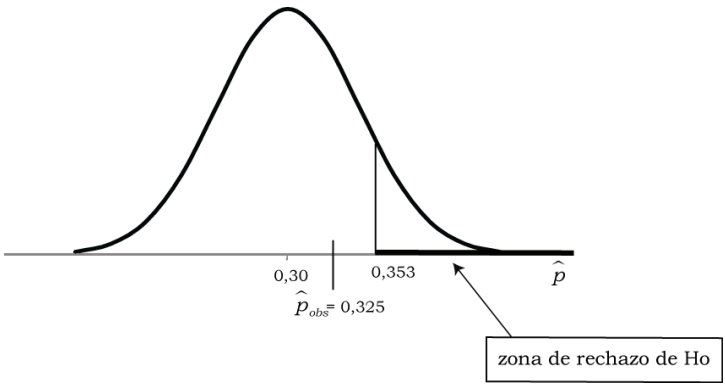
Para repetir la prueba sobre valores del estimador, vamos a transformar el punto crítico de acuerdo a la expresión general:

$$\hat{p} = P \pm z_c * \sqrt{\frac{P * (1 - P)}{n}}$$

de la que usaremos ambos signos cuando se trate de una prueba bilateral o solo la suma si es unilateral derecha o solo la resta si es unilateral izquierda. Para nuestro problema, corresponde sumar, por lo que resulta:

$$\hat{p} = P + z_c * \sqrt{\frac{P * (1 - P)}{n}} = 0{,}30 + 1{,}64 * \sqrt{\frac{0{,}03 * (1 - 0{,}30)}{200}} = 0{,}353$$

Representamos gráficamente la zona de rechazo como la cola derecha de la distribución de las $\hat{p}$:



El valor hallado en la muestra ($\hat{p}_{obs} = 0{,}325$) es menor que el punto crítico, por lo que no está en la zona de rechazo, no rechazamos la H_0 y concluimos que no hay evidencia para creer que el candidato haya aumentado su proporción de votos. Como había sucedido antes, la conclusión es la misma si trabajamos sobre los puntajes estandarizados o sobre valores del estimador.

Ejemplo 10.8

En el estudio Latinobarómetro[79] se menciona la importancia de comprender la percepción de la gente sobre el cambio

[79] Latinobarómetro es un estudio de opinión pública que aplica anualmente alrededor de 19.000 entrevistas en 18 países de América Latina representando a más de 400 millones de habitantes.

generacional, cómo vivían los padres en relación a ellos y cómo vivirán los hijos con respecto a ellos (expectativa futura). En el año 2004, el 58,2% de los argentinos encuestados consideraba que los hijos vivirían mejor que ellos. En el año 2005, los investigadores indagan nuevamente la expectativa futura. Dado que habían observado fluctuaciones en años previos se preguntan si esa expectativa positiva habrá cambiado o no respecto del año anterior. Las hipótesis que se formulan son:

$$H_0: P = 0{,}582$$
$$H_1: P \neq 0{,}582$$

En el estudio del año 2005, observaron que 676 participantes de los 1200 que componían la muestra consideraban que sus hijos vivirían mejor que ellos, siendo la proporción en la muestra:

$$\hat{p} = \frac{676}{1200} = 0{,}563$$

El nivel de significación se establece al 5%, por lo que los puntos críticos en puntajes z son $z_c = \pm 1{,}96$ (prueba bilateral).
Para responder a la hipótesis planteada, se transforma el valor de proporción observado a puntaje z haciendo:

$$z_{obs} = \frac{\hat{p}_{obs} - P}{\sqrt{\frac{P * (1-P)}{n}}} = \frac{0{,}563 - 0{,}582}{\sqrt{\frac{0{,}582 * (1-0{,}582)}{1200}}} = -1{,}36$$

El valor $z_{obs} = -1{,}36$ se encuentra entre los puntos críticos ±1,96, es decir que se halla en la zona de no rechazo de la H_0, por lo cual se concluye que no hay evidencia que indique que la proporción de expectativas positivas haya cambiado, la proporción no es significativamente distinta de 0,582.
Se alcanza la misma conclusión si se realizan los cálculos en función de los valores del estimador. Establecido el nivel de significación al 5%, se obtienen los puntos críticos en puntajes originales ($\hat{p}$) a partir de los puntos críticos en puntajes z (z_c):

$$\hat{p} = P \pm z_c * \sqrt{\frac{P * (1-P)}{n}} = 0{,}582 \pm 1{,}96 * \sqrt{\frac{0{,}582 * (1-0{,}582)}{1200}}$$
$$= 0{,}582 \pm 0{,}027$$

De esta manera se obtienen los $\hat{p}_c$ que delimitan las zonas de rechazo y no rechazo: 0,55 y 0,61. El valor $\hat{p}_{obs} = 0{,}563$ se

---

Corporación Latinobarómetro es una ONG sin fines de lucro con sede en Santiago de Chile, única responsable de la producción y publicación de los datos. http://www.latinobarometro.org/

encuentra comprendido entre ellos, por lo que no se rechaza la H_0.

Ejemplo 10.9
Otro de los aspectos indagados en el estudio Latinobarómetro se refiere al progreso en la reducción de la corrupción en las instituciones del Estado. En el año 2004, el 3,33% de los encuestados argentinos consideraba que se había progresado mucho en ese aspecto. En el estudio del 2005, se espera que los resultados sean más favorables debido a la aplicación de una serie de medidas tendientes a controlar la corrupción institucional. Así, las hipótesis que se formulan son:

$$H_0: P = 0{,}033$$
$$H_1: P > 0{,}033$$

De los 1200 argentinos encuestados en el año 2005, 51 señalan que se ha progresado mucho en reducir la corrupción institucional, es decir que la proporción en la muestra es

$$\hat{p} = \frac{51}{1200} = 0{,}043$$

El nivel de significación se establece en el 5%, dado que la prueba es unilateral derecha el único punto crítico en puntaje z es 1,64.
Luego, se transforma el valor de proporción observado en puntaje z, usando el estadístico de prueba:

$$z_{obs} = \frac{\hat{p}_{obs} - P}{\sqrt{\frac{P * (1-P)}{n}}} = \frac{0{,}043 - 0{,}033}{\sqrt{\frac{0{,}033 * (1-0{,}033)}{1200}}} = 1{,}94$$

El valor $z_{obs} = 1{,}94$ resulta superior al $z_c = 1{,}64$, por lo que se rechaza la H_0. Por eso concluimos que la proporción de personas que declaran que se ha progresado en la reducción de la corrupción institucional se incrementó de manera significativa.
Se obtiene la misma conclusión a partir de los valores del estimador. Establecido el nivel de significación al 5%, se obtiene el punto crítico en puntaje original ($\hat{p}_c$) a partir del punto crítico en puntajes z (z_c):

$$\hat{p} = P + z_c * \sqrt{\frac{P * (1-P)}{n}} = 0{,}033 + 1{,}64 * \sqrt{\frac{0{,}033 * (1-0{,}033)}{1200}}$$
$$= 0{,}033 + 0{,}008 = 0{,}041$$

El valor observado de $\hat{p}_{obs} = 0{,}043$ resulta mayor a $\hat{p}_c = 0{,}041$, por lo que se rechaza la H_0, y se concluye que la proporción de quienes creen que se ha progresado en la reducción de la

corrupción institucional se ha incrementado de manera significativa.

Ejemplo 10.10
Uno de los aspectos indagados en relación a la política, es el interés en la misma. En el estudio del año 2004, el 11% de los encuestados argentinos manifestó estar muy interesado en la política. Si bien no ha habido notables cambios en la última década con respecto al interés de los ciudadanos en la política, una serie de indicadores llevaron a los investigadores a considerar que el interés en esta cuestión podría haber disminuido. Las hipótesis planteadas son:

$$H_0: P = 0{,}11$$
$$H_1: P < 0{,}11$$

De la muestra de 1200 argentinos encuestados en el año 2005, 111 manifiestan un elevado interés en la política, la proporción en la muestra es:

$$\hat{p} = \frac{111}{1200} = 0{,}093$$

El nivel de significación se establece en el 5%, dado que la prueba es unilateral izquierda el punto crítico en puntaje z es -1,64.
Con el estadístico de prueba se transforma el valor de proporción observado en puntaje z:

$$z_{obs} = \frac{\hat{p}_{obs} - P}{\sqrt{\frac{P * (1-P)}{n}}} = \frac{0{,}093 - 0{,}11}{\sqrt{\frac{0{,}11 * (1-0{,}11)}{1200}}} = -1{,}88$$

El valor z_{obs} = *-1,88* resulta inferior al z_c = *-1,64*, por lo que se rechaza la H_0. Concluimos que la proporción de personas que manifiestan elevado interés en la política ha disminuido de manera significativa.
Como sucedía antes, podemos alcanzar la misma conclusión realizando los cálculos con el valor crítico del estimador:

$$\hat{p} = P - z_c * \sqrt{\frac{P * (1-P)}{n}} = 0{,}11 - 1{,}64 * \sqrt{\frac{0{,}11 * (1-0{,}11)}{1200}}$$
$$= 0{,}11 - 0{,}015 = 0{,}095$$

El valor observado de $\hat{p}_{obs} = 0{,}093$ resulta menor a $\hat{p}_c = 0{,}095$, por lo que se rechaza la H_0. Se concluye entonces que la proporción

de habitantes argentinos con alto interés en la política en el año 2005 es significativamente menor a 0,11.

Tipos de error en las pruebas de hipótesis

Dado que la decisión de aceptar o rechazar la H_0 se toma de manera probabilística, siempre existe la posibilidad de tomar una decisión incorrecta. Esto sucede porque las muestras son tomadas al azar y puede suceder que la que usamos para tomar la decisión sea una muestra extrema. Aunque es un resultado poco probable, no es imposible.

Como hemos visto, el nivel de significación mide la probabilidad de hallar un determinado resultado muestral si la H_0 fuera cierta, es una probabilidad pequeña, que habitualmente fijamos en 0,05 ó 0,01. Si la H_0 es cierta y la muestra sobre la que basamos la decisión es extrema, es decir, tiene un valor ubicado en alguna de las colas de la distribución, nuestra decisión será la de rechazar H_0 y esa decisión será errónea. Al momento de decidir, no podemos saber si H_0 es verdadera y obtuvimos una de esas muestras muy poco probables, o si efectivamente H_0 es falsa. Por esta razón el nivel de significación mide la probabilidad de errar en la decisión de esta manera: rechazando una H_0 que es verdadera. Éste se conoce como **Error de Tipo I** (ETI).

El **Error de Tipo I** es tomar una decisión errónea que consiste en rechazar la H_0 cuando esta es verdadera. Su probabilidad está fijada de antemano y es α, el nivel de significación de la prueba.

En consecuencia, establecer α es afirmar que se está dispuesto a correr ese riesgo de cometer el ETI. En un experimento que consiste en decidir si una droga produce efectos sobre un determinada patología, la Ho dirá que no hay efecto, por lo que cometer el ETI será creer que hay efecto (rechazar H_0) cuando en realidad no lo haya (H_0 verdadera). Como no sabemos si H_0 es verdadera o falsa, cada vez que rechacemos H_0 debemos recordar que hay una probabilidad α de haber tomado una decisión incorrecta. Esta incertidumbre está siempre presente en evaluación psicológica y educativa.

Veamos un ejemplo aproximado, pero familiar: las preguntas de un examen oral son una muestra de lo que el alumno sabe, si se usa un bolillero, la elección del tema que debe desarrollar es aleatoria. Supongamos que un alumno ha estudiado muy poco, pero la unidad que le toca desarrollar es alguna de las (muy

pocas) que sabe. En ese caso responderá correctamente y la decisión será que apruebe el examen. Si supiéramos que el alumno ignora todos los demás temas de la materia, la decisión correcta sería que no apruebe. La formalización de este problema es la siguiente: la hipótesis nula es la conservadora, esto es, que el alumno no sabe; será necesario sumar evidencia para que se tome la decisión de rechazar esa hipótesis y dar el examen por aprobado. La muestra de información que tiene el docente a su disposición (las bolillas que salieron por azar) es correctamente desarrollada por el alumno, lo que conduce a la decisión de rechazar la H_0 y dar el examen por aprobado. Para quien tiene toda la información, se ha cometido un Error de Tipo I, pero el docente nunca lo sabrá. En investigación nunca tenemos "toda la información", trabajamos con muestras, por lo que nunca sabemos si, cuando rechazamos H_0, estamos cometiendo este error o no.

Esta es la razón por la que α se elije con valores pequeños, pero no se puede reducir indefinidamente el valor de α, porque también existe el riesgo de aceptar H_0 siendo falsa.

Éste es otro tipo de error, al que llamaremos **Error de Tipo II** (ETII) y sucederá cuando aceptemos H_0 siendo falsa. En el experimento anterior, cometer este error consiste en creer que la droga no es efectiva (aceptar H_0) cuando en realidad sí tiene efectos (H_0 es falsa). En el ejemplo del examen, se trata de aplazar a un alumno (aceptar H_0) cuándo sí sabía, porque le tocó —a la inversa que en el caso anterior— una de las únicas bolillas que no sabía.

Se llama **Error de Tipo II** a la decisión equivocada de aceptar una hipótesis nula cuando ésta es falsa.

Según la prueba de que se trate, el costo de cometer cada tipo de error es diferente. Si se trata de evaluar el efecto de una intervención terapéutica, la H_0 dirá que no produce efectos. Entonces, cometer un ETI equivaldrá a recomendar la intervención y que no produzca efecto. Mientras que el ETII consistirá en desestimar una intervención que sí tenía efectos. Si se trata de una intervención muy riesgosa, el ETI es muy grave, porque implicará poner en peligro al paciente, por nada. Cometer ETII conlleva la pérdida del beneficio que la intervención habría dado. La decisión sobre qué error es más grave debe tomarse en cada caso y no pertenece al terreno de la estadística.

Ya hemos mencionado el error de una prueba diagnóstica consistente en que dé un resultado positivo al ser aplicada a alguien sano y llamamos a eso *falso positivo*. De modo más general, esa expresión indica que el error consiste en creer que "sucedió algo" cuando no fue así y corresponde, en las pruebas de hipótesis, al ETI.
De modo equivalente, hablamos del resultado negativo de una prueba diagnóstica al ser aplicada a alguien que sí está enfermo y lo llamamos *falso negativo*. Se trata del error opuesto, porque cuando se comete se cree que "no sucedió nada" cuando en realidad sí sucedió. Es el ETII.

Si insistimos sobre estos errores es para llamar la atención sobre dos aspectos fundamentales de las pruebas de hipótesis:
- Las conclusiones son probabilísticas, no son verdaderas ni falsas.
- Toda conclusión proveniente de estos procedimientos está sujeta a error.

ETII: μ, α y n
A diferencia del ETI, la probabilidad de cometer un ETII no es fijada de antemano, al contrario, el riesgo de creer que H_0 es verdadera cuando no lo es, depende de cuál sea la verdadera. Intuitivamente: no es igualmente probable creer que no hay diferencia entre dos valores cuando en realidad éstos son muy cercanos, que cuando difieren mucho, es más fácil confundir cosas que están cerca (creyendo que son iguales) que cuando están lejos.
Llamaremos β a la probabilidad de cometer el ETII y veremos cómo calcularla según las diferentes posibilidades del verdadero valor del parámetro sobre el cual se realiza la prueba.
El gráfico 9 corresponde a una prueba de hipótesis unilateral derecha sobre la media poblacional[80], con la forma:

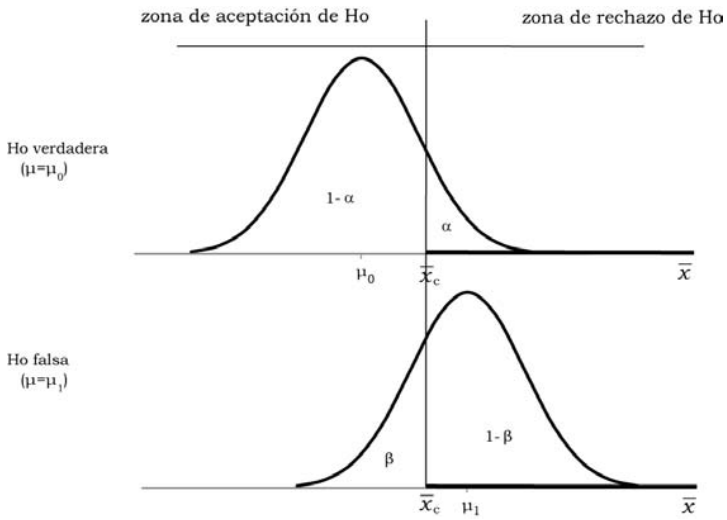
$$H_0: \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

Donde μ_0 es un valor determinado e hipotético para la media de la población. En el gráfico se ha ubicado ése valor hipotético y la zona de rechazo de H_0, estableciendo el nivel de significación (α) en 0,05. Además, hemos dibujado otra curva, centrada en otra

[80] Aunque ejemplificaremos solo para el caso de la media y para una prueba unilateral derecha, el concepto de errores de tipo I y II es general y vale del mismo modo para la proporción y para pruebas de hipótesis sobre otros parámetros, en pruebas unilaterales o bilaterales.

media poblacional que podría ser verdadera. Así, la curva superior representa la distribución de probabilidad de $\bar{x}$ si H_0 fuera verdadera ($\mu = \mu_0$). La curva inferior muestra la distribución de probabilidades de $\bar{x}$ si H_0 fuera falsa ($\mu > \mu_0$) y la verdadera media poblacional fuera μ_1 (que es una de las posibles entre las mayores que μ_0).

Gráfico 9: Comparación de la probabilidad de hallar a $\bar{x}$ en zona de aceptación o rechazo, según sea H_0 verdadera o falsa ($\alpha = 0{,}05$).



La parte superior del gráfico muestra que, si H_0 es verdadera, $\bar{x}$ tiene una probabilidad α de estar en la zona de rechazo, por lo que, si H_0 es verdadera y $\bar{x}$ está en esa zona, tomaremos la decisión errada de rechazarla. El complemento del nivel de significación (1-α), es la probabilidad de tomar la decisión correcta de aceptar H_0 cuando ésta es verdadera. Es la probabilidad de hallar a $\bar{x}$ en la zona de aceptación de H_0.
En la parte inferior del gráfico inferior vemos que, si H_0 es falsa, $\bar{x}$ tiene probabilidad 1-β de estar en la zona de rechazo, lo que llevará a tomar la decisión correcta de rechazar una H_0 falsa. Bajo el mismo supuesto de H_0 falsa, β es el área correspondiente a la zona de aceptación de H_0, por lo que mide la probabilidad de errar aceptando una H_0 falsa.
Debido a que no podemos tener certeza acerca de la verdad o falsedad de H_0, es que se nos plantean dos posibles escenarios: que H_0 sea verdadera o que sea falsa. En base a los datos con

que contamos en la muestra podemos tomar dos decisiones: aceptar H_0 o rechazarla, con lo que la decisión que tomemos puede ser correcta o incorrecta.

Esquema 1: Posibles combinaciones de estados de realidad, y decisión que se toma:

		Decisión	
		Aceptar H_0	Rechazar H_0
Estado de realidad	H_0 verdadera	Decisión correcta. Probabilidad: 1-α	Error de tipo I. Probabilidad: α
	H_0 Falsa	Error de tipo II. Probabilidad: β	Decisión correcta. Probabilidad: 1-β

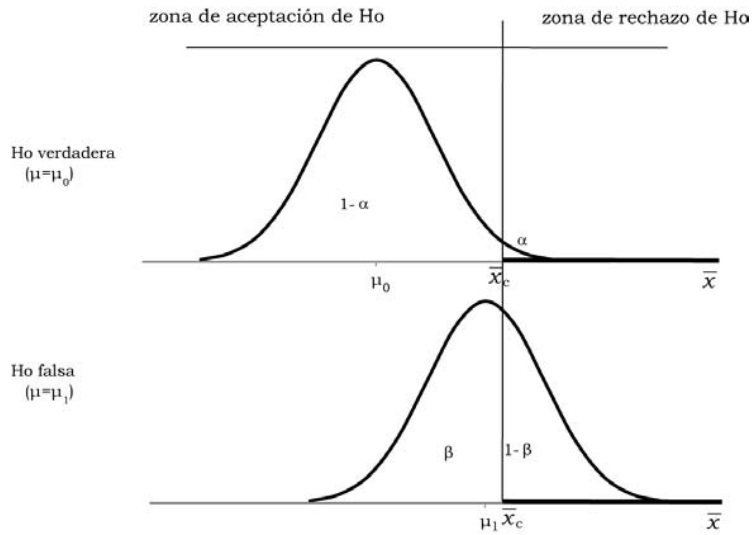
La última celda del esquema corresponde a la decisión correcta de rechazar una H_0 que es falsa, se denomina **potencia de la prueba** y es un indicador de la capacidad de la prueba para detectar hipótesis nulas que son falsas y rechazarlas.

Se llama **potencia de una prueba** a la probabilidad de rechazar una H_0 cuando ésta es falsa.

Es una importante medida de la calidad de la prueba, luego volveremos sobre su cálculo.

Así entonces, α es elegido por el investigador y mide el riesgo que está dispuesto a correr de rechazar una H_0 que es verdadera. Por el contrario, β depende de varios elementos. En primer lugar, depende de α: si se reduce el nivel de significación, aumenta el riesgo de cometer ETII. Si cambiamos el gráfico 9, reduciendo el nivel de significación, ahora la posición relativa de las áreas de rechazo y no rechazo queda:

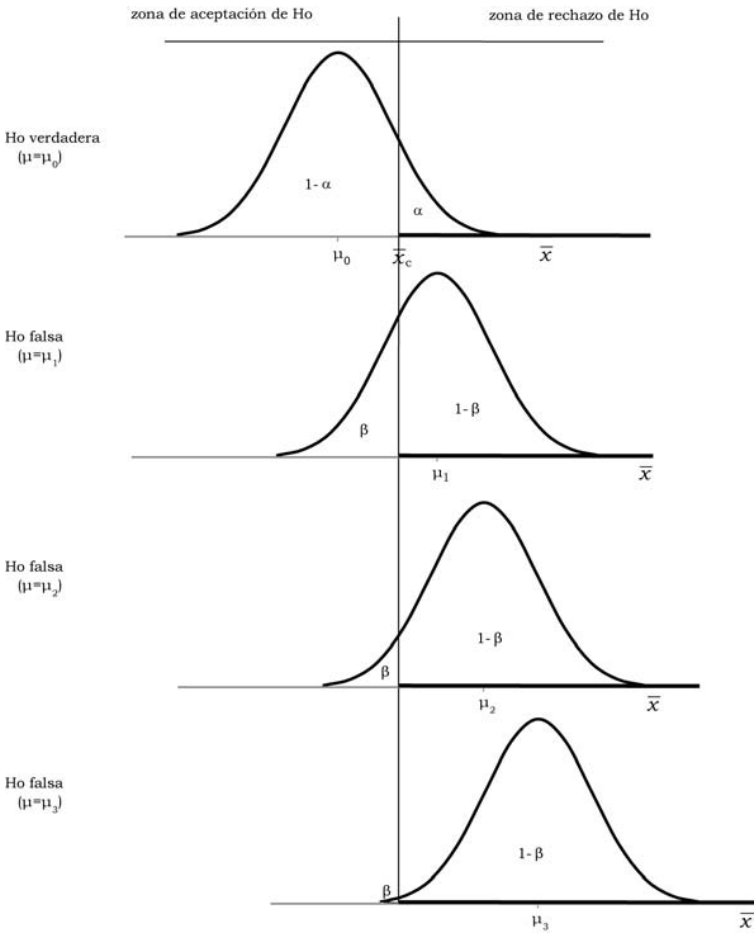
Gráfico 10: Comparación de la probabilidad de hallar a $\bar{x}$ en zona de aceptación o rechazo, según sea H_0 verdadera o falsa ($\alpha = 0{,}01$).



Como vemos, la reducción del nivel de significación del 5% al 1% hace que el punto crítico se desplace hacia la derecha y, en consecuencia, que aumente el área bajo la otra curva, que corresponde a H_0 falsa. Este cambio consiste en hacer a la prueba más exigente, al reducir las chances de rechazar H_0 por error del 5 al 1%. Su consecuencia es la de aumentar las chances de aceptar H_0 por error, aumentando β.

El ETII depende también de cuál sea la verdadera media poblacional. En los gráficos 9 y 10 planteamos como "otra posibilidad" que la verdadera media fuera μ_1, que es una de las formas en que puede ser H_0 falsa. Siendo H_0 falsa, μ puede tener distintos valores y ellos incidirán en la probabilidad de cometer ETII. En el gráfico siguiente, además de μ_1 agregamos otras dos medias poblacionales posibles μ_2 y μ_3.

Gráfico 11: Comparación de la probabilidad de hallar a $\bar{x}$ en zona de aceptación o rechazo, según sea H_0 verdadera o falsa, de tres modos diferentes.



El gráfico muestra que si la verdadera media poblacional difiere mucho de la hipotética (como es el caso de μ_3), es menor la probabilidad β de cometer ETII: β va decreciendo a medida que se consideran medias más alejadas de la que sostiene la H_0. Ésta es una manera de formalizar la idea intuitiva que mencionamos más arriba: es más fácil aceptar un valor equivocado de μ si el verdadero se le parece, que si es muy diferente. Más concreto aún: es más fácil aceptar por error un billete falso si se le parece mucho al verdadero que si es muy distinto, cuanto más difiera, menor será la probabilidad de aceptarlo por error.

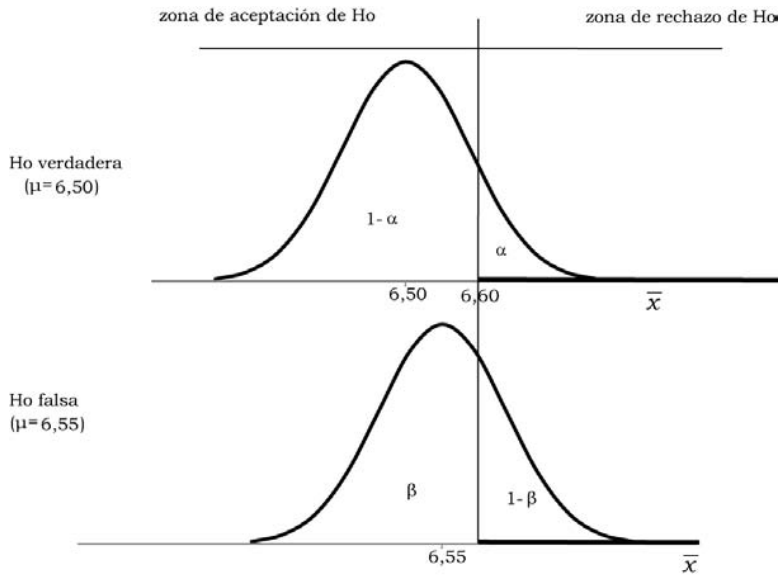
El gráfico 11 también muestra que, de manera complementaria, 1-β (la potencia de la prueba), va creciendo a medida que se consideran medias alternativas más alejadas de la hipotética.

Volvamos al ejemplo (ficticio) sobre el promedio con que egresan los estudiantes de Psicología en la prueba unilateral derecha, cuyas hipótesis son:

$$H_0: \mu = 6{,}50$$
$$H_1: \mu > 6{,}50$$

Preguntamos: si la verdadera media de nota con que egresan los estudiantes fuera de 6,55, ¿cuál habría sido la probabilidad de haber aceptado H_0? Dicho de otra forma ¿qué probabilidad hay de creer que la media sigue siendo 6,50 si en realidad ha aumentado a 6,55? Se trata de calcular la probabilidad de cometer ETII, porque se trata de aceptar una H_0 que es falsa. A un nivel de significación del 5%, el punto crítico que habíamos encontrado es 6,60, por lo que:

Gráfico 12: Ubicación de los tipos de error de la prueba si H_0 es verdadera o si es falsa.



La probabilidad de ETII en este caso es el área bajo la curva inferior que está por debajo del punto crítico, 6,60. Para calcular β es necesario hallar esa área bajo la curva normal centrada en 6,55, lo que requiere que se lo transforme a puntaje *z*:

$$z = \frac{6{,}60 - 6{,}55}{\frac{0{,}60}{\sqrt{100}}} = \frac{0{,}05}{0{,}06} = 0{,}83$$

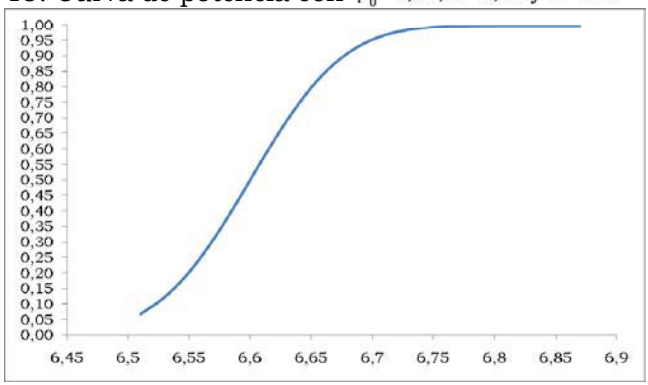
cuya área izquierda asociada (que buscamos con InfoStat®) es: P(z<0,83)=0,7977
Éste es el valor de β para esta prueba y esta media alternativa. Leemos el resultado diciendo que "si el verdadero promedio con que egresan los estudiantes de Psicología fuera de 6,55, habría una probabilidad de 0,7977 de creer que sigue siendo igual al histórico, de 6,50".
Calculamos la potencia de esta prueba, como el complemento de β:
Potencia=1-β =1-0,7977=0,2023

Curva de potencia
Dado que no se conoce la verdadera media de la población, solo podemos conjeturar acerca de ella y calcular el ETII, así como la potencia para diferentes valores posibles, como hemos hecho para tres valores en el gráfico 10. En general, resulta más valioso calcular la potencia (1-β) porque nos informa sobre la calidad de la prueba. El cálculo es largo porque deben seguirse los mismos pasos que en el ejemplo anterior para cada una de las medias posibles, pero podemos automatizarlo en una hoja de cálculo. La tabla para calcular la potencia de la prueba anterior se construye manteniendo el punto crítico fijo en 6,60 (y también la desviación estándar y el tamaño de la muestra), y calculando el puntaje z correspondiente a ese valor "desde" diferentes medias poblacionales alternativas, a las que hemos llamado μ_k. Así resulta:

Media alternativa μ_k	$z = \frac{6{,}60 - \mu_k}{\frac{0{,}60}{\sqrt{100}}}$	Probabilidad de ETII β	Potencia $1 - \beta$
6,51	1,50	0,9332	0,0668
6,53	1,17	0,8783	0,1217
6,55	0,83	0,7977	0,2023
6,57	0,50	0,6915	0,3085
6,59	0,17	0,5662	0,4338
6,61	-0,17	0,4338	0,5662
6,63	-0,50	0,3085	0,6915
6,65	-0,83	0,2023	0,7977
6,67	-1,17	0,1217	0,8783
6,69	-1,50	0,0668	0,9332
6,71	-1,83	0,0334	0,9666
6,73	-2,17	0,0151	0,9849
6,75	-2,50	0,0062	0,9938
6,77	-2,83	0,0023	0,9977
6,79	-3,17	0,0008	0,9992
6,81	-3,50	0,0002	0,9998
6,83	-3,83	0,0001	0,9999
6,85	-4,17	0,0000	1,0000
6,87	-4,50	0,0000	1,0000

Tenemos así diferentes resultados de la potencia, uno para cada media posible y vemos, ahora numéricamente, como disminuye el riesgo de aceptar H_0 a medida que la verdadera media es más lejana. Al mismo tiempo, aumenta la probabilidad de rechazar H_0 (la potencia). Esos resultados se representan gráficamente en lo que se denomina **curva de potencia**. Para el ejemplo anterior, la curva de potencia es la siguiente:

Gráfico 13: Curva de potencia con μ_0=6,50, α=0,05 y n=100



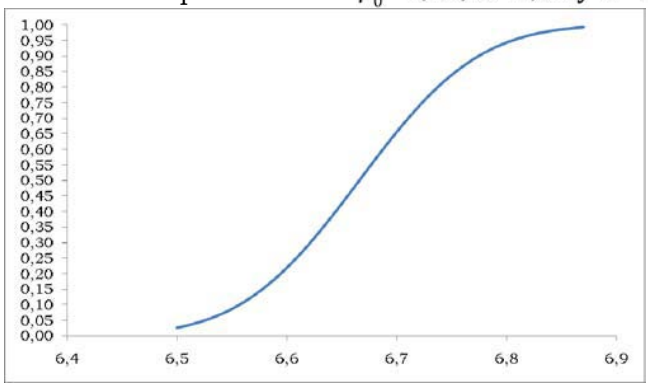
En el eje horizontal se ubican las posibles medias poblacionales (las distintas μ_k) y en el vertical, la potencia asociada a cada una de ellas. Vemos que si la verdadera media es la hipotética (μ=6,50), la probabilidad de rechazarla es α, y esa probabilidad aumenta a medida que se consideran medias más lejanas a . Si la verdadera media poblacional fuera tan lejana como 6,90, la probabilidad de rechazar H_0 es casi 1, lo que indica que es casi seguro que se rechazará esa hipótesis.

Dijimos antes que la potencia es una medida de la calidad de la prueba, en efecto, cuanto más rápidamente crezca esta curva, tanto más sensible será la prueba, porque será más probable detectar hipótesis nulas que son falsas y rechazarlas. Dicho de otro modo, será alta la probabilidad de rechazar una hipótesis nula si la verdadera media difiere —aunque sea poco—, de la hipotética.

La forma de la curva de potencia depende de varios factores, de los que solo nos detendremos en el tamaño de la muestra: con muestras más grandes, si todo lo demás se mantiene sin cambios, la potencia de la prueba aumenta. Inversamente, muestras pequeñas reducen la potencia.

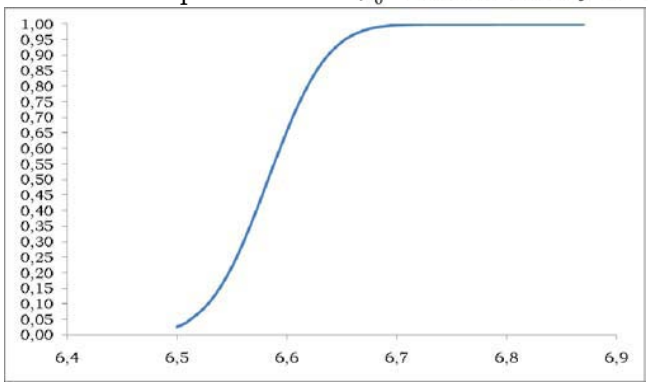
Veamos esto gráficamente, si la muestra fuera de menor tamaño que la del ejemplo, como *n=50*, la curva de potencia toma la siguiente forma:

Gráfico 14: Curva de potencia con μ_0=6,50, α=0,05 y n=50



Igual que antes en 6,50, la probabilidad de rechazo es 0,05, que es α. Pero cuando consideramos valores más lejanos de la media poblacional, la curva sube con más lentitud que la anterior, podríamos decir que "tarda más en reaccionar", son necesarios alejamientos más grandes para que crezca la probabilidad de rechazo.
Por el contrario, si se trata de una muestra de mayor tamaño, como 200 casos, la curva tiene la forma:

Gráfico 15: Curva de potencia con μ_0=6,50, α=0,05 y n=200



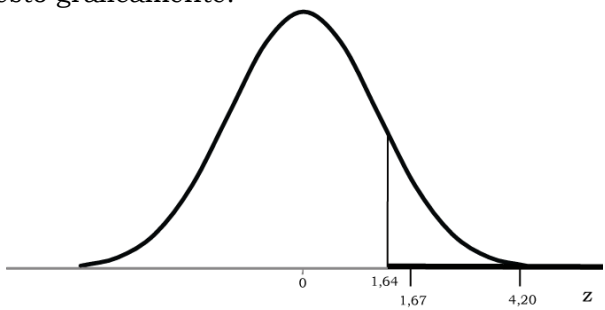
que muestra un aumento más rápido de la probabilidad de rechazar una H_0 cuando ésta es falsa.

Es importante destacar que la potencia no depende del resultado observado en la muestra, por el contrario, es un indicador de la calidad de la prueba como tal, independientemente de lo que se encuentre en la muestra.

Significación estadística y valor *p*

Hasta este punto vimos que el investigador establece el nivel de significación estadística al realizar una prueba de hipótesis, a partir de lo cual se determinan zonas de rechazo y no rechazo de la hipótesis nula y que esto puede hacerse sobre valores de *z* o del estimador. Luego, una vez obtenido el resultado de la muestra, se indica en qué zona se ubica, y se decide si se acepta o rechaza la H_0. La forma en que se comunican los resultados es, por ejemplo: "se rechaza la H_0 a un nivel del 5%". De modo que, como hemos dicho, se establece una regla de decisión y en base a ella se decide por sí o por no.

En una prueba unilateral derecha al 5%, el puntaje *z* que determina el punto de corte entre las zonas de aceptación y rechazo (*z* crítico) es 1,64, por lo que, si se obtiene un *z* observado de 1,67, se rechaza H_0 y también se rechaza si se obtiene 4,20, porque ambos son superiores al punto crítico. Sin embargo, encontrar *z=4,20* indica que el valor observado se aleja en gran magnitud del hipotético, mientras que si *z=1,67*, se rechaza muy cerca del límite. Entre dos situaciones en las que se rechaza H_0 puede haber diferencias de importancia; necesitamos una medida de la magnitud de la diferencia. Veamos esto gráficamente:



Aunque los dos valores de *z* llevan a concluir que se debe rechazar la H_0 (porque ambos están en la zona de rechazo), el valor 4,20 es sustancialmente menos probable que el 1,67, por lo que podríamos decir que en ese caso ($z_{obs}=4,20$) tenemos *más* evidencia para rechazar H_0, sin embargo, si solo informamos que "se rechaza H_0 a un nivel del 5%", quien lee no puede saber si se trató de una diferencia grande o pequeña.

El modo en que puede transmitirse esta información es ofreciendo el **valor de probabilidad** asociado al resultado muestral, que también se llama **valor *p***. Se trata de la probabilidad de observar en la experiencia un resultado igual o más extremo que el obtenido a partir de los datos muestrales,

bajo el supuesto de que la hipótesis nula es cierta. Es decir, la probabilidad de hallar un resultado como el que se encontró o más extremo que él, solo por azar. Un valor p pequeño indica que el resultado observado (o resultados más extremos que él) son poco probables bajo el supuesto de hipótesis nula cierta, por lo cual hay evidencia en contra de la hipótesis nula.
El valor p es una probabilidad condicional a la que escribimos formalmente como:

$$P\left(|u| \geq u_{obs} / H_0 V\right)$$

donde u es el estimador del parámetro al que se refiere la H_0. La expresión $|u| \geq u_{obs}$ es la forma reducida de decir que u sea más extrema (por encima o por debajo) que el valor observado.

Cuanto más pequeño es el valor p, tanta más evidencia hay para rechazar H_0. Por el contrario, un valor p grande indica que el resultado observado es muy probable bajo el supuesto de hipótesis nula cierta, lo que no aporta evidencia en contra de la hipótesis nula y conduce a no rechazarla.
Esta manera de indicar cuán esperable sería lo que hemos observado si H_0 fuera cierta, puede vincularse fácilmente con nuestro razonamiento anterior, comparando el valor p con el nivel de significación fijado de antemano.
Así resultan dos posibilidades:
Si el valor p es menor que el nivel de significación establecido ($p<\alpha$): se rechaza la hipótesis nula y se describe como "un resultado estadísticamente significativo". Esto quiere decir que la probabilidad de haber hallado este resultado por azar es pequeña, por lo que se trata de un efecto o diferencia que muy difícilmente se puede atribuir al azar.
Si el valor p es mayor que el nivel de significación establecido ($p>\alpha$): no se rechaza la hipótesis nula y se expresa como "un resultado no estadísticamente significativo". Lo que expresa que la probabilidad de haber hallado este resultado por azar es mayor que el máximo establecido, por lo que es razonable atribuirlo al azar, es decir a la variabilidad propia de los resultados muestrales.

Pero aun cuando esta modalidad de presentar el resultado puede hacerse equivalente al procedimiento anterior, nos aporta información adicional de importancia.

Veamos el modo de calcular estas probabilidades en los ejemplos que hemos tratado. Empezaremos por pruebas unilaterales:

En el ejemplo 10.3, promedio con que egresan los alumnos:

$$H_0: \mu = 6{,}50$$
$$H_1: \mu > 6{,}50$$

En la muestra de 100 alumnos habíamos hallado $\bar{x}_{obs} = 6{,}65$ y $s = 0{,}60$, con esos datos debemos calcular la probabilidad de hallar "un valor extremo como el observado o más extremo que él si H_0 es verdadera". Se trata de una probabilidad condicional, en la que la condición es que H_0 sea verdadera, lo escribimos así:

$$P(\bar{x} \geq 6{,}65 / \mu = 6{,}50)$$

que expresa la probabilidad de coincidir con 6,65 o superarlo (se trata de una prueba unilateral derecha, por eso usamos el signo ≥), si la media poblacional fuera la hipotética (6,50). Para encontrar la probabilidad, transformamos $\bar{x}$ a puntaje z:

$$z = \frac{6{,}65 - 6{,}50}{\frac{0{,}60}{\sqrt{100}}} = \frac{0{,}15}{0{,}06} = 2{,}50$$

Entonces debemos hallar $P(z \geq 2{,}50)$ y ya no hace falta indicar la condición (si $\mu = 6{,}50$), porque está incluida en el cálculo de z. Usando la opción de InfoStat®, obtenemos:

$$P(z \geq 2{,}50) = 0{,}0062$$

Ese es el valor p, e indica cuál es la probabilidad de encontrar un valor como el observado —o uno más extremo que él— si la H_0 fuera verdadera. Vemos que se trata de un valor pequeño, es menor a 0,05 y también menor a 0,01, por lo que podemos decir que la H_0 se rechaza a un nivel del 5% y también del 1%. Pero además de esto, comunicamos el valor p obtenido, porque eso da al lector una idea más completa de nuestro resultado.

Calculamos ahora el valor p asociado a la prueba del ejemplo 10.4 (del candidato):

$$H_0: P = 0{,}30$$
$$H_1: P > 0{,}30$$

En una muestra de $n = 200$ casos habíamos hallado $\hat{p}_{obs} = 0{,}325$. Por tratarse nuevamente de una prueba unilateral derecha, necesitamos conocer la probabilidad condicional de hallar un

valor como el observado —o más extremo que éste—si fuera verdadera H_0

$$P(\hat{p} \geq 0{,}325/P = 0{,}30)$$

Para calcular esta probabilidad transformamos el valor de $\hat{p}$ a puntaje z:

$$z = \frac{\hat{p} - P}{\sqrt{\frac{P * (1-P)}{n}}} = \frac{0{,}325 - 0{,}30}{\sqrt{\frac{0{,}30 * (1-0{,}30)}{200}}} = 0{,}77$$

Usando la hoja de cálculo encontramos que $P(z \geq 0{,}77) = 0{,}2206$. Se trata de una probabilidad "grande", si se compara con los niveles de significación que usamos, es superior a 0,05, por lo que la decisión será la de no rechazar H_0 y concluir que el candidato no mejoró su proporción de votos.

Muestras pequeñas y pruebas *t*

Cuando tratamos con distribuciones de probabilidad, en el capítulo 6, mencionamos la distribución *t de Student*, como un modelo de probabilidades con una forma similar a la distribución normal, pero más aplanada y dependiente de un dato al que denominamos "grados de libertad". Estos grados de libertad dependen del número de casos que haya en la muestra. Allí señalamos que a medida que los grados de libertad aumentan, la curva *t* tiende a asemejarse más a la normal. De modo que si se trata de muestras grandes, la distribución *t* es muy similar a la normal.

Vamos a usar esa distribución en las pruebas de hipótesis cuando trabajemos con muestras pequeñas, pero con una restricción. En aquellos casos en que sea posible suponer que la variable de origen (en la población) tiene distribución normal, entonces la distribución de las medias muestrales es adecuadamente modelada por la distribución *t*. Dicho de otro modo: si la variable tiene –en la población— distribución normal, entonces las medias muestrales, cuando se trate de muestras pequeñas, tiene *distribución t*. Nos encontramos así con distintas situaciones, de acuerdo al tamaño de la muestra y a la normalidad o no de la variable en la población.

Muestras grandes: Por el teorema central del límite, estamos autorizados a usar distribución normal para la media muestral, sin importar cuál sea la distribución de la variable en la población.

Muestras pequeñas: Si la variable tiene distribución normal en la población, usamos distribución *t de Student*, cuyos grados de libertad son *n-1*.
De estas dos condiciones (tamaño de muestra y normalidad de la variable en la población) resultan cuatro combinaciones:

		distribución de la variable en la población	
		Normal	no normal
tamaño de la muestra	grande $n \geq 30$	distribución normal, que equivale a *t de Student*	por TCL tiende a normal, que equivale a *t de Student*
	pequeña $n < 30$	distribución *t de Student*	ninguna[81]

De modo que hay una situación que no podemos abordar con estos procedimientos: la de muestras pequeñas provenientes de poblaciones no normales, para ellas disponemos de procedimientos llamados "no paramétricos" de los que nos ocuparemos en el capítulo 13. Las demás combinaciones pueden todas ellas resolverse con la distribución *t* que es específica en el caso de muestras pequeñas y es equivalente a la normal cuando la muestra es grande (porque lo son los grados de libertad).
Por esta razón, en la mayoría de los paquetes informáticos de análisis de datos, se habla de *pruebas t* de manera genérica para las pruebas sobre la media, y el estadístico de prueba constituye un valor *t* en lugar de *z*. La lógica es exactamente la que hemos seguido en este capítulo, solo que el programa hace las cuentas usando la distribución *t* que, cuando se trabaja con una muestra grande, da el mismo resultado que la normal.

Además de operar internamente con la distribución *t de Student*, los programas de análisis de datos ofrecen los resultados de las pruebas de hipótesis siempre en términos del valor *p*. Eso

[81] Existe un procedimiento basado en un resultado importante de la estadística que se conoce como "desigualdad de Tchebycheff", que establece que la probabilidad de encontrar a una variable a una distancia de su media superior a k desviaciones estándar es menor a $\frac{1}{k^2}$. Se expresa como $P(|x - E(x)| > k\sigma) < \frac{1}{k^2}$. También es posible recurrir a la distribución empírica, obtenida por remuestreo, InfoStat® lo hace en la opción *estimación por bootstrap* para construir intervalos de confianza.

permite que quien lee la salida pueda decidir si rechaza o no rechaza la H_0, de acuerdo al nivel de significación que haya fijado.

La siguiente salida InfoStat® corresponde a una prueba de hipótesis acerca del promedio del número de hijos por mujer. Se busca evidencia para decidir si, según los datos de una muestra de 116 casos, la población de la que esa muestra proviene tiene en promedio un hijo por mujer o si ese promedio es menor a 2. Las hipótesis entonces son:

$$H_0: \mu = 2$$
$$H_0: \mu < 2$$

Prueba t para una media

Valor de la media bajo la hipótesis nula: 2

Variable	n	Media	DE	T	p(Unilateral)
hijos	116	1,98	1,22	-0,18	0,4326

En el primer renglón, la salida ofrece el valor de μ bajo H_0, que es 2. Luego muestra el nombre de la variable (hijos), la cantidad de casos en la muestra, la media muestral ($\bar{x} = 1{,}98$), la desviación estándar muestral (DE=1,22), el valor del estadístico de prueba (t=-0,18) y el valor de probabilidad asociado (p(Unilateral I)=0,4326). Esta última probabilidad especifica que se trata de una prueba unilateral izquierda (rechaza para valores menores que el hipotético).
La media muestral tiene un valor que está muy levemente por debajo del hipotético (1,98 frente a 2), por lo que podríamos suponer que la diferencia no es significativa, sin embargo, con la apreciación subjetiva no es suficiente, es necesario observar el valor p de la salida. Esta última cifra es grande (43,26%) por lo que no se rechaza la H_0. Concluimos que no hay evidencia para considerar que la población de la que proviene la muestra tenga un nivel de fecundidad inferior a dos hijos por mujer.

Resumen de las pruebas sobre una muestra mencionadas en el capítulo

Parámetro	Estimador	Estadístico de prueba	Puntos críticos en términos del estimador	Supuestos
μ	$\bar{x}$	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$	$\bar{x}_c = \mu \pm z * \frac{s}{\sqrt{n}}$	$n \geq 30$ ó distribución normal de x en la población
P	$\hat{p}$	$z = \frac{\hat{p} - P}{\sqrt{\frac{P * (1 - P)}{n}}}$	$\hat{p}_c = P \pm x * \sqrt{\frac{P * (1 - P)}{n}}$	$n \geq 100$

Usando el valor p, la lectura del resultado de la prueba de hipótesis se expresa:

Si la hipótesis nula fuera verdadera, habría una probabilidad p de hallar un valor como el observado o uno más extremo.

Actividad práctica de repaso 10

1. Los resultados que se muestran a continuación provienen de una muestra de alumnos que cursaron Psicoestadística en 2009. El promedio histórico que resulta del primer parcial según registros de años anteriores, es de 6,66. Considere la siguiente salida:

```
Prueba T para un parámetro

Valor del parámetro probado: 6,66
Variable      n      Media   DE     T     p(Bilateral)
Primero       303     6,82  2,13   1,30         0,1957
```

a. ¿Cuál es la unidad de análisis?
b. ¿Qué variable se analiza?
c. ¿Cuántos casos se consideran?
d. ¿Cuánto vale el promedio muestral?
e. ¿Cuál es la hipótesis nula de la prueba?
f. ¿Cuál es la hipótesis alternativa?
g. ¿Cuántos son los grados de libertad de la prueba?
h. ¿Cuál es la lectura del valor p?
i. Si el nivel de significación es del 5%, ¿cuál es la conclusión?

2. En la siguiente salida se solicitó el intervalo de confianza:

```
Prueba T para un parámetro

Valor del parámetro probado: 6,66
Variable      n      Media   DE    LI(95) LS(95)  T  p(Bilateral)
Primero       303     6,82  2,13     6,58   7,06 1,30      0,1957
```

Complete lo siguiente:
a. Si la media poblacional fuera…. habría una probabilidad de… % de...

3. Cuando se cambia la confianza se obtiene:

```
Prueba T para un parámetro

Valor del parámetro probado: 6,66
Variable      n      Media   DE    LI(99) LS(99)  T  p(Bilateral)
Primero       303     6,82  2,13     6,50   7,14 1,30      0,1957
```

Complete lo siguiente:
a. Si la media poblacional fuera…. habría una probabilidad de… % de...

4. Ahora nos interesa conocer si la proporción de aprobados difiere del valor histórico que, según los registros es de 85%. Para ello, recodificamos las notas asignando el valor cero (0) a los menores a cuatro y uno (1) a los cuatro y superiores. La variable dicotómica que así resulta se llama aprobó y se puede tratar con el mismo procedimiento que una cuantitativa. La salida InfoStat ® para esta prueba es:

```
Prueba T para un parámetro

Valor del parámetro probado: 0,85
Variable      n      Media   DE     T     p(Bilateral)
aprobo        303     0,91  0,29   3,46         0,0006
```

a. ¿Qué es la "media", cuyo valor es 0,91?
b. ¿Cuál es la hipótesis nula de la prueba?
c. ¿Cuál es la hipótesis alternativa?
d. ¿Cuál es la lectura del valor p?
e. Si el nivel de significación es del 5%, ¿cuál es la conclusión?

5. Al pedir también el intervalo de confianza, obtenemos:

```
Prueba T para un parámetro

Valor del parámetro probado: 0,85
Variable      n      Media   DE    LI(95) LS(95)  T  p(Bilateral)
aprobo        303     0,91  0,29     0,87   0,94 3,46      0,0006
```

Complete:
a. Si la proporción poblacional fuera…. habría una probabilidad de… % de...

3. En una prueba estandarizada de atención, la media del número de aciertos es, para la población general, de 25. Se cree que las personas que pasan por un período de depresión, podrían alcanzar puntajes más bajos en esta prueba.
a. ¿Cuál es la H_0?
b. ¿Cuál es la H_1?
c. ¿Cuál es la lateralidad de la prueba?
d. ¿En qué consistiría cometer Error de Tipo I en esta prueba?
e. ¿En qué consistiría cometer Error de Tipo II en esta prueba?

Capítulo 11: Comparación entre dos grupos

Eduardo Bologna
Andrés Urrutia

En el capítulo anterior presentamos la prueba de hipótesis como un procedimiento para decidir si un valor observado en la muestra es compatible con el valor poblacional que plantea la hipótesis nula, para la media o la proporción. Trabajamos sobre una única población y pusimos a prueba un valor determinado para un parámetro (media o proporción). Ahora ampliamos nuestro análisis, porque vamos a poner a prueba la eventual diferencia entre dos grupos. Probaremos por ejemplo, si puede aceptarse que dos poblaciones tengan la misma media en una variable cuantitativa o también la misma proporción de casos en una categoría de una variable cualitativa. Como antes, la hipótesis nula será la de no diferencia, es decir que formularemos como H_0 que las dos poblaciones tienen la misma media o bien la misma proporción.

En este capítulo, abordaremos los procedimientos que permiten comparar las medias o proporciones de dos grupos y determinar si las diferencias que se encuentran son significativamente mayores de lo que pudiera esperarse por puro azar. Es decir, recorreremos los pasos necesarios para tomar una decisión en términos estadísticos, a favor o en contra de la hipótesis que sostiene que dos grupos son iguales respecto del parámetro bajo análisis. Un modo alternativo de expresar el problema es considerar que lo que se pone a prueba es si las dos muestras provienen de la misma población o de dos poblaciones diferentes.

En el diseño experimental —tema que se verá en detalle en Metodología de la Investigación—, a menudo interesa conocer, si un grupo sometido a cierto tratamiento muestra cambios diferentes que los que experimenta otro grupo que no fue sometido a ese tratamiento. Los procedimientos que veremos se utilizan para determinar si las ganancias obtenidas en una muestra de sujetos tras un tratamiento —por ejemplo de entrenamiento cognitivo—, son lo suficientemente amplias como para representar diferencias en la población. En este tipo de casos, realizamos una primera evaluación del grupo (pre-test), obtenemos la media de ciertas variables (aciertos en una prueba de atención, palabras recordadas, aciertos en la

correspondencia entre nombres y caras, etc.). Luego sometemos a los sujetos a un período de entrenamiento y a continuación los evaluamos nuevamente (post-test). Necesitamos determinar si los cambios, medidos a través de la diferencia de los promedios, son de una magnitud tal que podamos atribuirlos al entrenamiento o bien si pueden explicarse por azar.
Los procedimientos destinados a comparar grupos tienen gran difusión en investigación. Muchos de los problemas que interesa resolver usándolos se encuentran vinculados a distintos campos de la Psicología y la Educación, ya que se comparan grupos que han recibido una droga con quienes no la han recibido, que han pasado por un período de entrenamiento o no, que han participado en grupos terapéuticos o han desarrollado otra actividad, alumnos que aprenden con uno u otro método, etc. En todos los casos se realiza una comparación en algún resultado, que puede ser el puntaje en un test o cualquier otra variable sobre la que se busca intervenir. Por ejemplo, en una tesis de Maestría en Psicología Clínica, el autor separa a un grupo de niños deficientes mentales en dos grupos. Uno de ellos participa de un taller con actividades ecológicas (se lo llama "grupo experimental"), mientras que el otro grupo no lo hace (es el denominado "grupo control"). A través de pruebas estandarizadas se mide la sociabilidad, esperando que las actividades en las que participaron pudieran estimularla. Se comparan los niveles de sociabilidad de los niños que participaron del taller con los de quienes no participaron, esta evaluación se hace antes y después de la realización de estas actividades, que se prolongaron por un mes. En este ejemplo hay dos comparaciones: la socialización de los niños que participaron con la de los que no lo hicieron, y la de la sociabilidad de los que participaron antes del taller y luego de él. En ambos casos se utilizan las pruebas sobre las que trata este capítulo.

Siempre la hipótesis nula sostendrá que no hay diferencias entre las puntuaciones obtenidas antes y después del tratamiento, o por un grupo y el otro; o lo que es lo mismo, que las diferencias son cero (nulas). En la hipótesis alternativa se planteará la situación opuesta, sea bilateral (las medias difieren) o unilateral (una media es mayor que la otra).
Sin embargo estos procedimientos no sólo se utilizan en experimentos, sino también en estudios que analizan datos de encuestas o de observación. Así una variable independiente puede constituirse por grupos o muestras de sujetos que conforman categorías diferentes. Varones versus mujeres, población rural versus población urbana, niños que repiten o no repiten un grado. Pueden

interesarnos entonces, ciertas diferencias en los promedios de alguna variable entre estos grupos. ¿Tienen los mismos ingresos promedio los varones jefes de hogar que las mujeres que son jefas de hogar? ¿La proporción de analfabetos es la misma entre provincias del NOA o del Centro del país? La edad promedio de las madres primerizas que se atienden en hospitales públicos, ¿es la misma que la de quienes van a privados? Podemos comparar varones y mujeres y evaluar si hay diferencia entre el promedio de un grupo y otro en relación a la introversión, el neuroticismo o su afabilidad. También podríamos comparar personas mayores y jóvenes en relación a los promedios que obtuvieran en pruebas de inteligencia fluida o cristalizada. En estos casos, como investigadores esperamos que ciertas puntuaciones promedios sean más altas en los adultos jóvenes (inteligencia fluida) versus los promedios que pudieran obtener los mayores, probablemente más altos en los desempeños de pruebas de inteligencia cristalizada. En estos ejemplo, jefes y jefas de hogar, NOA y Centro del país, jóvenes y viejos, mujeres y varones, etc. son grupos independientes, que se diferencian por los valores de una variable (posición dentro del hogar, región, edad, género) y nos interesa investigar si estos agrupamientos evidencian diferencias en otras variables de interés: como los ingresos, la proporción de analfabetos, la inteligencia, la introversión.

Cuando apreciamos que las medias de dos grupos son diferentes, no podemos saber a priori si las muestras provienen de poblaciones cuyas medias son diferentes o no, porque las diferencias observadas entre los resultados muestrales pueden provenir de azar. Por eso, encontrar que $\bar{x}_1 \neq \bar{x}_2$ no nos conduce inmediatamente a que $\mu_1 \neq \mu_2$.

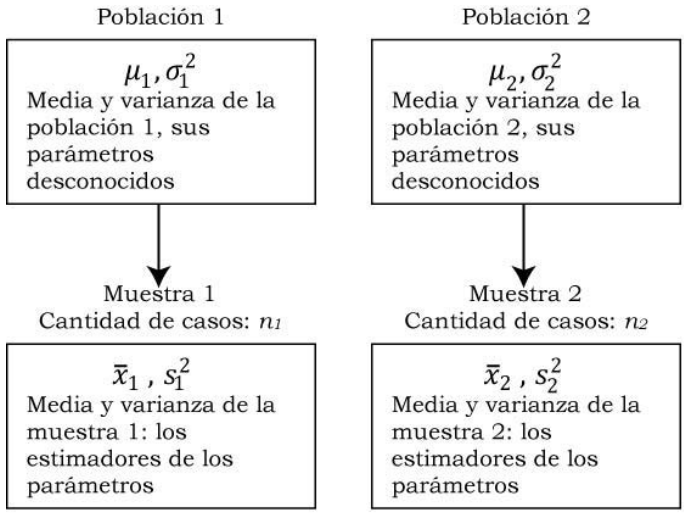
En un ejemplo elemental pero ilustrativo, seleccionamos aleatoriamente 100 estudiantes de Agronomía varones y preguntamos por su estatura, obtenemos una media de 1,73m. Luego repetimos la medición a otros 100 estudiantes varones (también aleatoriamente elegidos) de Ingeniería Civil y la media es de 1,75m. Aunque estas medias muestrales sean diferentes, no podemos concluir que, en conjunto, todos los estudiantes de Ingeniería Civil son más altos que los de Agronomía, porque esta pequeña diferencia muy probablemente se deba al azar que introduce el procedimiento de muestreo.

Nuevamente, nos encontramos con el problema de decidir cuándo una diferencia es lo suficientemente grande como para considerarla debida a una diferencia a nivel de las poblaciones, solo que ahora, en lugar de confrontar el resultado muestral con una media poblacional hipotética, comparamos las medias provenientes de dos poblaciones.

El procedimiento para comparar los parámetros (medias o proporciones) de dos poblaciones consiste en extraer una muestra de cada población: llamaremos a los tamaños de esas muestras n_1 y n_2. En cada muestra calcularemos —como antes—, los estimadores correspondientes. Si se estima una diferencia de medias calcularemos las medias ($\bar{x}_1$ y $\bar{x}_2$), así como las desviaciones estándar (s_1 y s_2) de cada una. Si lo que se estima es una diferencia de proporciones, en cada muestra se calculará la proporción de casos en la categoría de interés: $\hat{p}_1$ y $\hat{p}_2$ que son los estimadores de los parámetros P_1 y P_2
Veamos en primer lugar el caso en que la variable bajo análisis es cuantitativa y entonces nos interesa comparar las medias poblacionales. El parámetro que se estima es la diferencia entre las medias poblacionales: $\mu_1 - \mu_2$, y la estimación se hace a través de la diferencia entre las medias muestrales: $\bar{x}_1 - \bar{x}_2$. El error estándar de ese estimador (que será necesario para estandarizar el valor observado) va a depender del tipo de prueba de que se trate. En este capítulo trataremos dos situaciones, cuando sean muestras independientes, y muestras apareadas o dependientes.

Muestras independientes

Prueba de diferencia de medias
La comparación de las medias de dos poblaciones independientes se realiza comparando los correspondientes estimadores de las medias poblacionales, veamos el siguiente esquema para recordar la notación:



Una vez que se dispone de los datos muestrales, el estadístico de prueba toma la forma;

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

Es decir, se trata de la diferencia entre la diferencia de las medias muestrales y la diferencia de las medias poblacionales, dividida por el error estándar de la diferencia (el error estándar del estimador). La expresión mantiene la estructura que mencionamos en el capítulo anterior: estimador menos parámetro sobre el error estándar del estimador. En este caso el estimador es la diferencia de las medias muestrales y el parámetro es la diferencia de las medias poblacionales.

Este estadístico de prueba tiene distribución normal si los tamaños de las muestras son suficientemente grandes (mayores a 30 casos) y tiene distribución *t de Student* si se trata de muestras pequeñas y puede además suponerse que las variables que se analizan tienen distribución normal en la población. Debido a que la distribución *t* tiende a la normal a medida que aumenta el tamaño de la muestra, y como lo hicimos antes para una sola muestra, escribiremos de manera general:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

Como sabemos, cuando las muestras sean grandes, las probabilidades asociadas al valor de *t* coincidirán con las de la distribución normal.

El denominador del estadístico de prueba puede ser calculado de dos manera diferentes y eso va a depender de que podamos suponer que las varianzas de las dos poblaciones de las que provienen las muestras son iguales o que no sea así.

Caso 1: Varianzas poblacionales iguales $\sigma_1^2 = \sigma_2^2$

Si las varianzas poblacionales son iguales, entonces las dos varianzas que se calculan desde las muestras, constituyen dos estimadores del mismo parámetro. Cada muestra ofrece una estimación de la varianza, que es la misma en las dos poblaciones de origen. En ese caso, calcularemos primero un promedio[82] de las dos estimaciones

[82] La fórmula de este promedio tiene en cuenta que las varianzas provienen de muestras de diferente tamaño y que los denominadores de las varianzas son *n-1*. Por eso no es simplemente la suma de ambas dividida dos.

dadas por las varianzas muestrales, a la que llamaremos varianza combinada:

$$s_{comb}^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}$$

Usando este estimador de la (única) varianza poblacional, el estadístico de prueba asume la forma:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{comb} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Cuyos grados de libertad se calculan, en función de los tamaños de las muestras como: $gl = n_1 + n_2 - 2$.

Caso 2: Varianzas poblacionales diferentes $\sigma_1^2 \neq \sigma_2^2$

Si no es posible suponer que las varianzas de las poblaciones de donde provienen las muestras son iguales, entonces debemos usar las varianzas muestrales de manera separada. Cuando éste es el caso, el estadístico de prueba es:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Esta expresión es más sencilla, pero el cálculo de los grados de libertad de la distribución *t de Student* se vuelve más complejo. La fórmula para hacerlo es[83]

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

[83] Esta expresión se conoce como la ecuación de Welch–Satterthwaite y fue desarrollada en el campo de la teoría de errores, los programas de análisis de datos la aplican de manera automática cuando las varianzas son diferentes a cierto nivel de significación.

No vamos a usar esta expresión para calcular los grados de libertad, pero es la que usan los programas de análisis de datos cuando detectan que las varianzas poblacionales no son iguales.

De acuerdo a esto, cuando vamos a hacer una prueba t para comparar las medias de dos grupos debemos antes saber si estamos ante el caso 1 ó el caso 2; lo que significa que deberemos antes decidir si las varianzas de las dos poblaciones pueden considerarse iguales o no. Dado que no conocemos estas varianzas poblacionales, la decisión se toma a partir de los datos muestrales, es decir, a partir de las varianzas halladas en las muestras (s_1^2 y s_2^2). Se realiza una prueba, cuya hipótesis nula afirma que las varianzas poblacionales son iguales y su resultado permitirá decidir, a un determinado nivel de significación, si puede tratarse a las varianzas poblacionales como iguales o si debe considerárselas diferentes. No nos ocuparemos de esa prueba, pero es importante conocer la forma de este razonamiento, tanto para comprender la manera de solicitar el procedimiento al programa, como para interpretar el resultado.

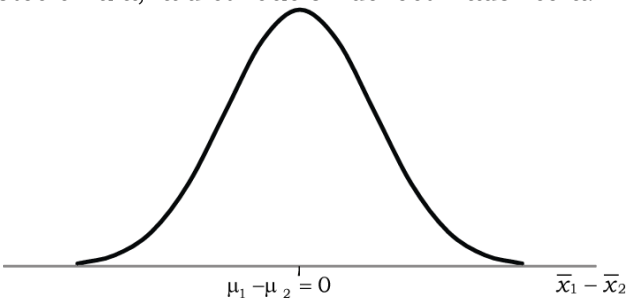
Ejemplo 11.1
Nos preguntamos si, para una carrera universitaria dada, el tiempo que tardan en completar sus estudios los estudiantes que trabajan y los que no trabajan es el mismo o si difiere. Para ello tomamos una muestra de 100 alumnos que trabajan y obtenemos una media de duración de la carrera de 6,7 años y una desviación estándar de 1,2 años. Extraemos otra muestra, de 150 casos de estudiantes que no trabajan, en la que obtenemos un promedio de los años de duración de 6,3 con una desviación estándar de 1,5 años. Todos estos datos provienen de la información descriptiva que proveen las muestras, son los que vamos a usar para hacer la inferencia acerca de las poblaciones, constituidas por el total de alumnos que trabajan y que no trabajan.
Como antes, la hipótesis nula es la del "no cambio", la que afirma que no hay diferencia, por lo que:

$$H_0: \mu_1 - \mu_2 = 0$$

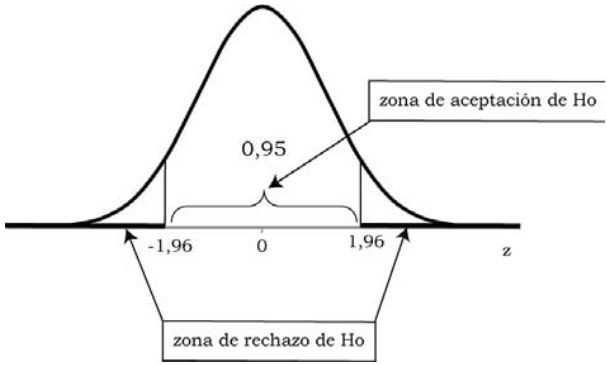
que dice que la diferencia entre las medias poblacionales es cero (es nula). Hemos dicho que nuestro interés está en saber si las medias poblacionales son iguales o si difieren, por lo que se trata de una prueba bilateral, entonces la hipótesis alternativa indicará que:

$$H_1: \mu_1 - \mu_2 \neq 0$$

Para la presentación continuaremos usando la distribución normal, ya que resulta más familiar, recordemos que como indicamos en el capítulo 11, los paquetes de análisis de datos usan directamente distribuciones *t*.
Bajo la hipótesis nula, la distribución del estimador será:



Para decirlo nuevamente, la variable aleatoria es la diferencia de medias muestrales, el centro de la distribución es el parámetro, que según la hipótesis nula es cero.
Fijamos el nivel de significación de la prueba en el 5% y los puntos críticos resultan, sobre la distribución normal estándar ±1,96, gráficamente:



Ahora transformamos a puntaje *z* los valores observados en las muestras. Recordemos para ello que el puntaje *z* se define como la diferencia entre el estimador y el parámetro, dividida por el error estándar del estimador. Para esta prueba, el parámetro es la diferencia de medias poblacionales, su estimador es la diferencia de

medias muestrales. Supongamos que las varianzas poblacionales pueden suponerse iguales (caso 1), entonces[84]:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{comb} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

en la que

$$s^2_{comb} = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}$$

Con los datos de este ejemplo, la varianza combinada es:

$$s^2_{comb} = \frac{(100 - 1) * 1{,}2^2 + (150 - 1) * 1{,}5^2}{100 + 150 - 2} = \frac{477{,}81}{249} = 1{,}92$$

por lo que la desviación estándar combinada resulta:

$$s_{comb} = \sqrt{1{,}92} = 1{,}38$$

Como la hipótesis nula establece que la diferencia de medias poblacionales es cero $\mu_1 - \mu_2 = 0$, entonces, el estadístico de prueba nos queda:

$$z = \frac{(6{,}7 - 6{,}3) - 0}{1{,}38 * \sqrt{\frac{1}{100} + \frac{1}{150}}} = \frac{0{,}4}{0{,}18} = 2{,}22$$

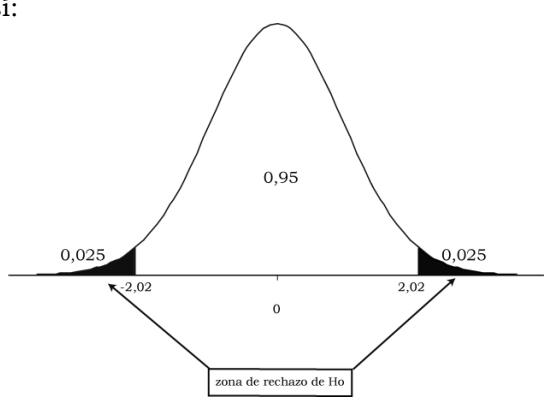
Este es el estadístico de prueba correspondiente a la diferencia de medias para muestras independientes cuando las varianzas poblacionales pueden suponerse iguales. Este valor de z (2,22) se encuentra en la zona de rechazo de Ho, por lo que la decisión será la de rechazar H_0 y concluir que el tiempo que tardan en terminar la carrera los alumnos que trabajan difiere significativamente del que tardan los que no trabajan. Dicho de otro modo: la diferencia observada en las medias de las muestras es significativa a un nivel del 5%.

[84] Cuando se solicita a InfoStat®, automáticamente (a menos que pidamos lo contrario) se verifica si puede suponerse que las varianzas poblacionales son iguales o no y, en base al resultado de esa verificación, se calcula el error estándar de la diferencia con una fórmula o con otra.

Si los mismos resultados se hubiesen hallado en muestras más pequeñas, por ejemplo de 20 y 25 casos de estudiantes que trabajan y no trabajan respectivamente, y además hubiésemos podido suponer que las dos poblaciones son normales en la variable *duración de la carrera*; entonces habría correspondido usar una prueba *t*[85]. Los grados de libertad de esta distribución se calculan sumando los tamaños de muestra y restando dos:

$$gl = n_1 + n_2 - 2 = 20 + 25 - 2 = 43$$

Con 43 grados de libertad, los valores *t* que delimitan un 5% extremo en la distribución de probabilidades son ±2,02, que gráficamente se representan así:



Para tomar la decisión calculamos el estadístico de prueba con el mismo procedimiento anterior, para lo que es necesario contar primero con la desviación estándar combinada. Calculamos primero la varianza combinada:

$$s^2_{comb} = \frac{(20-1) * 1{,}2^2 + (25-1) * 1{,}5^2}{20 + 25 - 2} = \frac{81{,}36}{43} = 1{,}89$$

y la desviación estándar combinada es:

$$s_{comb} = \sqrt{1{,}89} = 1{,}37$$

por lo que el estadístico de prueba resulta:

$$t = \frac{(6{,}7 - 6{,}3) - 0}{1{,}37 * \sqrt{\frac{1}{20} + \frac{1}{25}}} = \frac{0{,}4}{0{,}41} = 0{,}97$$

[85] Recordemos que la distribución normal vale para muestras grandes, por el Teorema Central del Límite.

Este valor no se ubica en la zona de rechazo de H_0, por lo que la conclusión será en este caso que no hay evidencia para creer que el tiempo promedio que tardan en terminar la carrera los alumnos que trabajan sea diferente que el que tardan quienes no trabajan.

La comparación de las conclusiones de estos dos ejemplos muestra algo muy importante: una misma diferencia absoluta de 0,4 años en los promedios muestrales de los grupos, es significativa cuando proviene de muestras grandes (100 y 150 casos) y deja de serlo cuando se obtiene en muestras pequeñas (20 y 25 casos). Como vimos en el capítulo anterior, el tamaño de las muestras incide en la potencia de la prueba, en su capacidad para detectar diferencias.

Por las razones que ya hemos mencionado, en el cálculo informatizado del estadístico de prueba no se distingue entre utilizar distribución normal (*z*) para muestras grandes y *t de Student* para muestras chicas, como acabamos de hacer en los ejemplos anteriores. Por el contrario, se usa siempre distribución *t*[86]. Así, esta prueba se conoce como *prueba t de diferencia de medias*. Es el caso del programa InfoStat® que estamos utilizando.

Ejemplo 11.2
La salida que mostramos a continuación es una comparación de la edad promedio de los docentes varones y mujeres del nivel medio. Los datos provienen de una muestra aleatoria de 246 docentes de la ciudad de Córdoba. La hipótesis nula de esa comparación es que no hay diferencia en la edad promedio de mujeres y varones docentes del nivel medio, que se confronta con una hipótesis alternativa que afirma que sí hay diferencia. Se trata de una prueba bilateral que se expresa así:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

donde los subíndices 1 y 2 se refieren a los grupos formados por docentes mujeres y varones respectivamente. La salida InfoStat® tiene la siguiente forma:

[86] Recordemos que esta distribución tiende a la normal cuando las muestras son grandes.

Prueba T para muestras Independientes

Clasific	Variable	Grupo(1)	Grupo(2)	n(1)	n(2)	media(1)	media(2)	P (Var.Hom.)	T	p	prueba
sexo	edad	{M}	{V}	180	66	40,42	40,71	0,5295	-0,23	0,8183	Bilateral

La primera columna (Clasific) indica cuál es la variable de clasificación, es decir la que define los grupos.
Luego se especifica la variable sobre la que se hará la comparación (la edad).
Los grupos (mujeres y varones).
La cantidad de casos en cada uno de ellos (180 mujeres y 66 varones).
Las dos columnas siguientes indican las medias muestrales: la edad promedio de las mujeres de la muestra es de 40,42 años y la de los varones 40,71.
La columna siguiente se usa para probar si las varianzas de los grupos son iguales o diferentes. P(Var.Hom) quiere decir la probabilidad asociada a la hipótesis de varianzas homogéneas (i.e. suficientemente parecidas como para tratarlas como iguales). Es el valor *p* que resulta de la prueba de igualdad de varianzas. En este caso es 0,5295, en un valor grande (mucho mayor a 0,05 que es el nivel de significación usual) por lo que no se rechaza la igualdad de las varianzas. Por lo tanto, lo que sigue de la prueba de diferencia de medias corresponde al caso 1, varianzas iguales. Esto permite decidir cuál es la fórmula para calcular el t_{obs}.
Luego encontramos el valor del estadístico de prueba *t*, es el t_{obs} para los datos del problema.
Sigue el valor de probabilidad asociado a ese t_{obs}, que es de 0,8183.
La última columna indica que se trata de una prueba bilateral que fue lo que solicitamos al programa.

Debido a que el valor de probabilidad es un número alto (81%), muy superior al 5% que solemos usar como criterio para rechazar H_0, la decisión es la de no rechazarla. Por lo que concluimos que no hay evidencia para creer que en las escuelas de nivel medio de la ciudad de Córdoba, la edad promedio de los docentes varones sea diferente de la de las mujeres.

La lateralidad de la prueba que analizamos en el capítulo anterior, sigue del mismo modo cuando trabajamos con diferencias de medias. Por ejemplo, podemos tener expectativa en que una droga que se está experimentando produzca efectos sobre la depresión, en dirección a

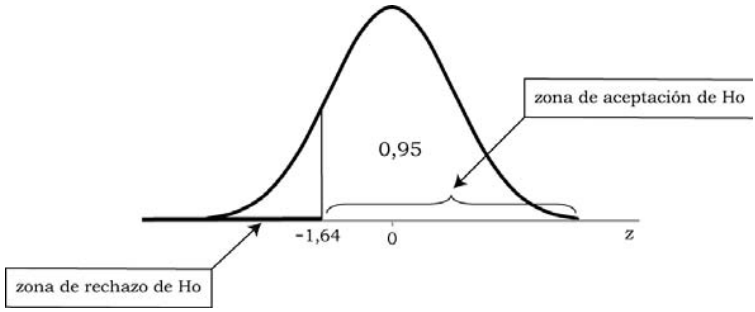
reducir el puntaje que alcanzan los sujetos en un test que la evalúa. Para poner a prueba la droga será necesario diseñar un experimento, no avanzaremos en ese tema pero de manera muy simple, podemos pensar en dos grupos de pacientes depresivos a uno de los cuales se administra la droga (grupo 1) y al otro no (grupo 2)[87]. Concluiremos que la droga tiene efectos si luego de un tiempo de su administración, el grupo de pacientes que la recibieron experimenta cambios positivos en mayor magnitud que el otro. Dicho de otro modo, habrá efectos si la media de puntaje en el test que evalúa depresión es significativamente menor en el grupo que se sometió al tratamiento. Como siempre, la hipótesis nula afirmará que no hay diferencia:

$$H_0: \mu_1 - \mu_2 = 0$$

Como ahora nos interesa que el grupo 1 haya reducido su puntaje en el test de depresión, esperamos que la media del grupo 1 sea menor que la del grupo 2, lo cual se escribe:

$$H_1: \mu_1 - \mu_2 < 0$$

Por la lateralidad de la prueba, solo hay una zona extrema de rechazo, la izquierda. Fijando un 5% de nivel de significación, si se tratara de muestras grandes, en las que podemos usar distribución normal, esta zona se representa como ya sabemos:



Por el contrario, si se trata de muestras pequeñas, corresponde usar distribución *t*, y el punto crítico dependerá de los grados de libertad.

Ejemplo 11.3
El grupo de 15 sujetos que recibió un medicamento arroja un puntaje promedio en la prueba que evalúa la depresión de 5,7 puntos, con

[87] Se trata de una versión muy simplificada, solo para ver el uso de esta prueba. El tema se desarrolla con detalles en la materia Metodología de la Investigación Psicológica.

desviación estándar de 1,1 puntos. Los 10 pacientes que no recibieron el medicamento alcanzan un puntaje promedio de 7,2 puntos, con desviación estándar de 1,6 puntos. La prueba previa sobre la homogeneidad de las varianzas indica que no puede suponérselas iguales, por lo que estamos en el caso 2. Usaremos distribución t con grados de libertad que deben calcularse con la expresión de Welch-Satterthwaite, el resultado[88] es 15. El punto crítico $t_{15;\ 0,05}$, es -1,76.

El estadístico de prueba vale:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} = \frac{(5,7 - 7,2) - 0}{\sqrt{\frac{1,1^2}{15} + \frac{1,6^2}{10}}} = -2,58$$

Es un valor que se sitúa a la izquierda del punto crítico, por lo que se rechaza la H_0 y se concluye que el grupo que recibió la droga alcanzó un puntaje significativamente menor que el otro grupo.

Ejemplo 11.4

En un estudio realizado en el Centro de Promoción del Adulto Mayor, se seleccionaron aleatoriamente entre todos los asistentes, a 503 personas mayores de 50 años. Entre otras preguntas, se computó el número de hijos de cada uno de ellos.

Se conformaron dos grupos con los sujetos estudiados, quienes tenían hasta 65 años en el momento del estudio y quienes tenían 66 ó más años. En este ejemplo, estos dos grupos, si bien no fueron asignados de manera aleatoria, se comportarían como muestras independientes.

Podríamos suponer que las personas de mayor edad tendrían más hijos, dadas ciertas razones históricas, religiosas y culturales que habitualmente inciden en el control de la natalidad. Esta hipótesis está basada también en el conocimiento demográfico; sabemos que las tasas de fecundidad en nuestro país han venido descendiendo desde hace varias décadas. Sostenemos así que los que conforman el grupo de 66 años y más tendrán más hijos que los de 65 años o menos, con lo que se trata de una prueba unilateral. Si llamamos 1 al grupo de los de hasta 65 años y 2 al de quienes tienen 66 ó más, las hipótesis se expresan de la siguiente manera:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 < 0$$

Los resultados descriptivos de los dos grupos se muestran a continuación:

[88] Redondeado al entero

Estadística descriptiva

edadcod	Variable	n	Media	D.E.
1,00	hijos	243	2,49	1,44
2,00	hijos	260	2,61	1,38

El nombre de la variable edadcod, corresponde a los grupos 1, hasta 65 años y 2, de 66 ó más. Los grupos tienen 243 y 260 personas respectivamente.
Vemos que hay diferencia en el número de hijos promedio de los dos grupos. Los del segundo grupo (66 ó más) tienen en promedio 0,12 hijos más que los del otro grupo. Ahora bien, dadas estas diferencias entre las medias encontradas, nos interesa determinar si se deben al azar o bien si tienen una magnitud tal que representen una diferencia que pueda atribuirse a la diferencia de edad de los sujetos de los dos grupos.
Planteamos una prueba *t* para muestras independientes, fijamos el nivel de significación en 5% y hallamos que el valor crítico de *t* con 501 grados de libertad (243+260-2) es -1,65 (como habíamos señalado, cuando los grados de libertad son elevados, casi no hay diferencia entre una distribución *t* y una normal). La región de rechazo está conformada por todos los valores de *t* que sean inferiores a -1,65.

InfoStat® presenta la salida de esta prueba de la siguiente manera:

Prueba T para muestras Independientes

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	pHomVar	T	p-valor	prueba
edadcod	hijos	{1.00}	{2.00}	243	260	2.49	2.61	0.5191	-0.94	0.1747	Unilateral

Que nos muestra: la variable que clasifica (edadcod), la que se analiza (hijos), los nombres de los dos grupos (1 y 2), los tamaños de muestra en cada grupo (243 y 260), las medias (2,49 y 2,61), el valor *p* para la prueba de homogeneidad de varianzas (0,5191), el valor del estadístico de prueba (t=-0,94), el valor de probabilidad asociado a la prueba y el tipo de prueba, unilateral en este caso.
La prueba preliminar de homogeneidad de varianzas da un valor *p* de 0,5191, que es mucho mayor a 0,05 que usamos como nivel de significación, por lo que no se rechaza la hipótesis nula de esa prueba: las varianzas no difieren significativamente. Por lo tanto, la prueba de diferencia de medias se hace considerando varianzas iguales y los grados de libertad se calculan directamente como n_1+n_2-2.
El t_{obs} de la prueba es -0,94, que se ubica en la zona de no rechazo (la zona de rechazo es la que queda por debajo de -1,65). Con ese

resultado ya estamos en condiciones de concluir sobre la prueba: no se rechaza H_0, las medias no difieren de manera significativa, no hay diferencia en el número de hijos entre los grupos conformados.
Si no hubiésemos calculado previamente el punto crítico, podemos llegar a esta conclusión a partir del valor *p*, que en este caso es 0,3495 y representa la probabilidad de haber hallado esa diferencia o una mayor, por puro azar. Este es el valor que, como sabemos, se juzga en comparación con el nivel de significación, por ser mayor a 0,05 que habíamos establecido, decidimos no rechazar la H_0. Cuanto más pequeña sea esta probabilidad, tanta más evidencia habrá para rechazar H_0, en este caso la consideramos grande y no rechazamos. La lectura del valor *p* es, en este caso "si el número promedio de hijos de los dos grupos fuera el mismo, la probabilidad de haber encontrado una diferencia de 0,12 ó superior, es de 0,3495". Brevemente:

$$P((\bar{x}_1 - \bar{x}_2 < -0{,}12)/(\mu_1 - \mu_2 = 0)) = 0{,}3495$$

Prueba de diferencia de proporciones
Compararemos ahora dos muestras en cuanto a la proporción de casos que hay en una categoría de una variable que puede ser nominal o también de un nivel superior[89]. Aunque en los programas de análisis de datos este procedimiento se realiza también solicitando *prueba t*, esto sólo está autorizado si se trabaja con muestras grandes. De manera muy similar a la prueba de diferencia de medias, la H_0 afirma que no hay diferencia entre las proporciones de las dos poblaciones, por lo que tiene la forma:

$$H_0: P_1 - P_2 = 0$$

mientras que la H_1 puede indicar que las proporciones solo difieren:

$$H_1: P_1 - P_2 \neq 0$$

Cuando la prueba es bilateral o bien, si es unilateral derecha:

$$H_1: P_1 - P_2 > 0$$

[89] Aunque la prueba se usa principalmente para comparar variables nominales, no hay inconveniente en usarla con variable métricas, definiendo con precisión qué se compara. Por ejemplo, si la variable es la edad, podemos comparar la proporción de personas mayores de 65 años en dos muestras. Todos los mayores de 65 constituyen una categoría y el resto la otra. En ese caso decimos que hemos "dicotomizado" una variable métrica, la hemos transformado en una variable con dos categorías: mayores de 65 y de 65 ó menos. Sin embargo este procedimiento implica una pérdida, porque no se usan las potencialidades de análisis que ofrecen las variables cuantitativas.

o izquierda:

$$H_1: P_1 - P_2 < 0$$

Dado que solo trabajaremos con muestras grandes al analizar la diferencia de proporciones, solo usaremos la distribución normal. El estadístico de prueba tiene forma similar al que usamos en la prueba de la proporción para una sola muestra, solo que ahora hay que hacer participar los datos que provienen de dos muestras. Así resulta:

$$z = \frac{(\hat{p}_1 - \hat{p}_1) - (P_1 - P_2)}{\sqrt{\frac{\hat{p}_1 * (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 * (1 - \hat{p}_2)}{n_2}}}$$

donde $\hat{p}_1$ y $\hat{p}_2$ son los estimadores de P_1 y P_2 respectivamente. Notemos que a diferencia de la prueba sobre una sola proporción, ahora no contamos con valores poblacionales de P_1 ni de P_2, ya que la H_0 solo enuncia que son iguales. Por eso ahora, el error estándar del estimador (el denominador del estadístico de prueba) se calcula en base a las proporciones muestrales.

Ejemplo 11.5
Para ilustrar el procedimiento de prueba de hipótesis sobre la comparación de proporciones entre dos muestras independientes, consideremos que disponemos de datos de la proporción de aplazos sobre el total de materias rendidas que tienen varones y mujeres estudiantes de una carrera universitaria. En una muestra de 200 mujeres se encuentra una proporción de 15% de aplazos, mientras que entre los 150 varones que componen la otra parte de la muestra, el porcentaje es del 17%. Nos preguntamos si se trata de una diferencia significativa o si puede explicarse por la variabilidad propia de los datos. Así, las hipótesis de la prueba serán:

$$H_0: P_1 - P_2 = 0$$
$$H_1: P_1 - P_2 \neq 0$$

Es decir que la hipótesis nula afirma que la diferencia en la proporción de aplazos entre varones y mujeres es cero en la población, y la hipótesis alternativa, dice que esa diferencia es diferente de cero.
Fijamos el nivel de significación en el 5% y, por tratarse de una prueba bilateral, los puntos críticos son ±1,96.
A continuación calculamos el estadístico de prueba, que resulta:

$$z = \frac{(\hat{p}_1 - \hat{p}_1) - (P_1 - P_2)}{\sqrt{\frac{\hat{p}_1 * (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 * (1 - \hat{p}_2)}{n_2}}} = \frac{(0{,}15 - 0{,}17) - 0}{\sqrt{\frac{0{,}15 * (1 - 0{,}15)}{200} + \frac{0{,}17 * (1 - 0{,}17)}{150}}}$$
$$= -0{,}50$$

que es un valor que no se ubica en la zona de rechazo, por lo que la decisión es la de aceptar H_0 y concluir que no hay evidencia para creer que la proporción de aplazos entre mujeres y varones sea diferente.

Muestras apareadas

Esta prueba es muy valiosa cuando se buscan diferencias que puedan haber aparecido en sujetos individuales, su nombre indica que las diferencias se consideran en pares o en parejas. En lugar de comparar muestras provenientes de dos poblaciones independientes, compararemos unidades de análisis consigo mismas, en dos momentos distintos. Ejemplo típico de este uso lo constituyen las pruebas en las que se compara una situación "antes y después". Así, cuando cierta característica de los sujetos experimentales se evalúa antes de un tratamiento y se vuelve a medir luego del tratamiento, interesa conocer en qué medida se han producido cambios *individuales*. Observamos a cada individuo en dos momentos y comparamos las dos observaciones (mediciones). En el contexto del diseño experimental, la medición en el momento inicial se llama *pretest* y la posterior *postest*.

Pero no es el único ámbito en que se usa, cuando se comparan los logros que los alumnos alcanzan luego de una determinada experiencia pedagógica, es valioso poder comparar el "estado inicial" (anterior a la intervención) con el "estado final" (posterior) de cada individuo, o cuánto aprendieron o si se modificó determinada conducta. Esta prueba pone el acento en los cambios sucedidos en cada sujeto y no en las diferencias entre sujetos. En medicina, por ejemplo, si interesa conocer la diferencia de presión arterial a la mañana y al atardecer, corresponde medir la presión arterial de cada persona en los dos horarios y comparar uno a uno. Por cierto, luego agregaremos el resultado, pero la diferencia se mide sujeto a sujeto.

No hay cambio conceptual en el planteo de las hipótesis: la H_0 afirma que no hay diferencia entre las dos mediciones (o entre las mediciones hechas en momentos distintos) y la H_1 podrá ser unilateral o bilateral. Sin embargo cambiamos levemente la notación, porque vamos a

resumir la diferencia en la letra $\overline{D}$ (por media de diferencias) y equivale a $\mu_1 - \mu_2$. El objetivo de este cambio es poner el acento en que no tratamos con dos poblaciones independientes (una representada por μ_1 y la otra por μ_2) sino con las diferencias entre dos mediciones. $\overline{D}$ es un parámetro y es sobre $\overline{D}$ que formulamos ahora las hipótesis:

$$H_0: \overline{D} = 0$$

La hipótesis alternativa puede ser bilateral

$$H_1: \overline{D} \neq 0$$

Unilateral derecha

$$H_1: \overline{D} > 0$$

O bien unilateral izquierda

$$H_1: \overline{D} < 0$$

El estimador de $\overline{D}$ se llamará $\bar{d}$ y es la media de las diferencias individuales, que tiene distribución *t de Student* con *n-1* grados de libertad. Como se trata de una sola variable, el estadístico de prueba es semejante al que usábamos para una sola muestra, porque resume las diferencias para cada caso:

$$t = \frac{\bar{d} - \overline{D}}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

La última igualdad se debe a que, bajo la H_0, $\overline{D} = 0$.
s_d es la desviación estándar de la variable d, que mide las diferencias individuales.

Ejemplo 11.6
Se trata de ejemplo incompleto desde el punto de vista del diseño[90], pero útil para mostrar esta técnica. Se busca detectar el eventual efecto que tendría una intervención terapéutica sobre el bienestar de pacientes diagnosticados de depresión. Disponemos de un test que nos permite evaluar el bienestar de los pacientes y que esa medición alcanza un nivel cuantitativo, en una escala que va de 0 (mínimo bienestar) a 10 (máximo bienestar). Para evaluar la terapia, nos interesamos por los cambios que suceden en el bienestar de los

[90] Como se verá en Metodología de la Investigación, este ejemplo carece de "grupo control", por lo que no puede asegurarse que los cambios observados se deban a la intervención terapéutica o a otros factores.

pacientes entre el momento previo y el posterior a la misma. Hacemos esto por medio de la aplicación del test a una muestra de 7 pacientes diagnosticados de depresión, en dos momentos: antes de la terapia y al cabo de ella.
Esperamos que la terapia tenga por efecto el de aumentar el bienestar de los sujetos que la reciben, dicho de otro modo, esperamos cambios positivos en el puntaje del test. Las hipótesis son entonces:

$$H_0: \overline{D} = 0$$
$$H_1: \overline{D} > 0$$

Fijado el nivel de significación en el 5% y suponiendo que la variable se distribuye de manera normal en la población, el punto crítico de la distribución t con 6 grados de libertad, de una cola derecha es t_c =+1,943.

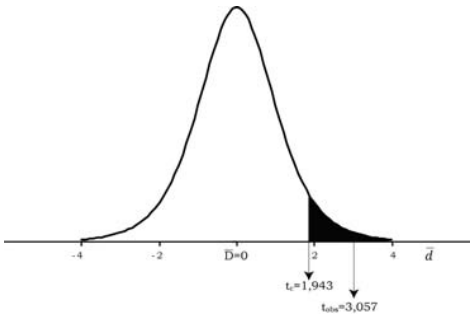
Los siguientes son los datos recogidos:

sujeto	medición previa x_1	medición posterior x_2	diferencia d_i
1	5	7	2
2	4	7	3
3	6	6	0
4	3	5	2
5	4	5	1
6	5	6	1
7	7	7	0

Indicamos en esta tabla (proveniente de la matriz de datos) a cada individuo y al puntaje que obtuvo en la prueba que mide bienestar tomada antes de la terapia (x_1) y en la tomada después (x_2). Además hemos calculado las diferencias para cada sujeto, restando la primera de la segunda, a esas diferencias las llamamos d_i, el subíndice i se refiere a cada sujeto y va desde 1 hasta 7 en este ejemplo.
La hipótesis nula asociada a esta prueba será que no hay diferencia en las dos mediciones, es decir que no hay efectos de la terapia, que el bienestar de los sujetos luego del tratamiento es igual que antes de él.
El cambio respecto de las pruebas para muestras independientes es que ahora pasamos a trabajar con d como variable, en ella calcularemos los estadísticos descriptivos: la media de las d_i es $\bar{d} = 1{,}286$, y su desviación estándar es $s_d = 1{,}113$.
Con esa información calculamos el estadístico de prueba:

$$t_{obs} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{1{,}286}{\frac{1{,}113}{\sqrt{7}}} = 3{,}057$$

Por ser mayor que el punto crítico ($t_{obs} > t_c$), este valor pertenece a la zona de rechazo de H_0, con lo que concluimos que hay evidencia para considerar que los puntajes en el test de bienestar efectivamente son mayores luego de la terapia. La siguiente es la representación gráfica de este resultado.



Solicitada a InfoStat®, la prueba ofrece el siguiente resultado:

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	T	p(Unilateral D)
x2	x1	7	1.29	1.11	3.06	0.0112

En la última columna, denominada "P(Unilateral D)" que es el tipo de prueba que solicitamos, nos agrega el valor de probabilidad asociado a *t*, que vale 0,0112. Se trata de una probabilidad pequeña (menor que el nivel de significación, fijado en 0,05) que nos confirma el rechazo de H_0.

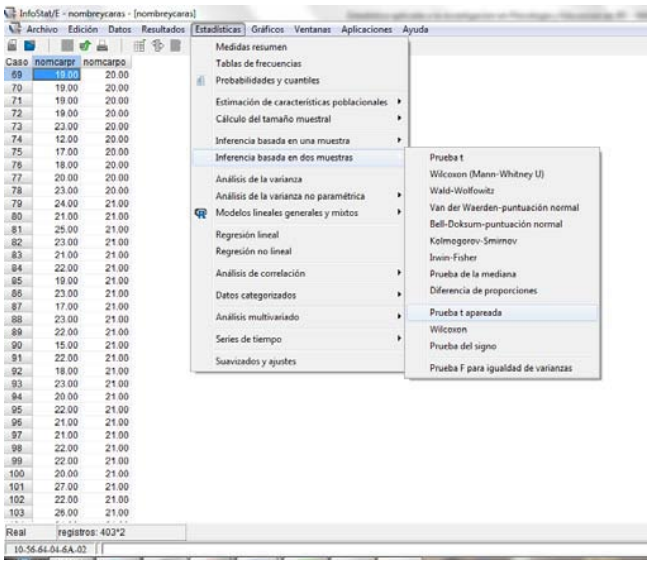
Ejemplo 11.7

Se trata de un estudio que se llevó a cabo con personas mayores que asisten a centros de día, programas universitarios y centros de jubilados. Los participantes de este estudio manifestaban quejas de atención y memoria, fundamentalmente a corto plazo. Las quejas están centradas en cierta falta de competencia en tareas domésticas: pérdida de llaves, dejar una hornalla prendida o ciertas ejecuciones cotidianas afectadas por un declive normal del funcionamiento cognitivo. Todos los sujetos fueron entrenados en destrezas básicas asociadas a la mejora de estas habilidades cognitivas, durante 12 sesiones grupales de dos horas duración, con una frecuencia

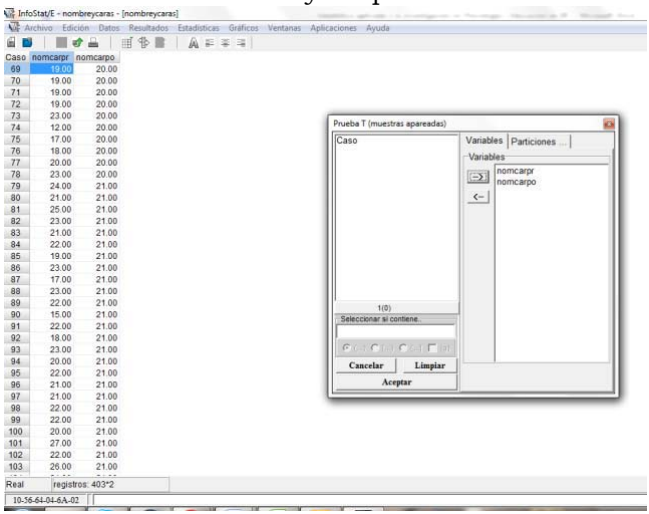
semanal. Basado en la teoría del procesamiento de la información, los alumnos mayores aprenden ciertas técnicas que les permiten controlar los automatismos motores, ser más detallistas, prestar más atención a información que se les escapa, y a través de la imaginación, del uso del lenguaje y de diversas técnicas, aprenden a mejorar su capacidad asociativa. Esta actividad se denomina "taller de memoria".

Los sujetos fueron evaluados con pruebas neuropsicológicas que valoran las habilidades entrenadas, al inicio y finalizada la intervención.

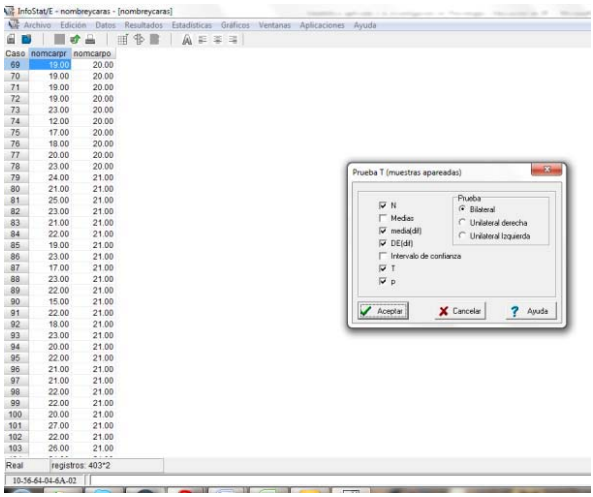
Una de las pruebas toma en cuenta los puntajes de una prueba de asociación entre caras y nombres, que consiste en presentar a las personas fotos y sus correspondientes nombres y apellidos. Luego de un tiempo, en el que deben aprenderse las caras con sus correspondientes nombres, se les presentan sólo las fotos y ellos deben completar los nombres y apellidos. Se le asigna un punto por cada nombre y por cada apellido correcto, sin descontar puntos si obtuvieron errores. La hipótesis nula dirá que no hay diferencia entre los puntajes obtenidos antes y después de la intervención, mientras que la hipótesis alternativa afirmará que sí hay diferencia. Tomados los puntajes de 401 sujetos, solicitamos a InfoStat® una prueba t apareada:



Luego indicamos las variables que dan los puntajes en la prueba aplicada antes del entrenamiento y después de él:



Las variables se llaman nomcarpr (nombres y caras pretest) y nomcarpo (nombres y caras postest). Luego de aceptar, tenemos algunas opciones:

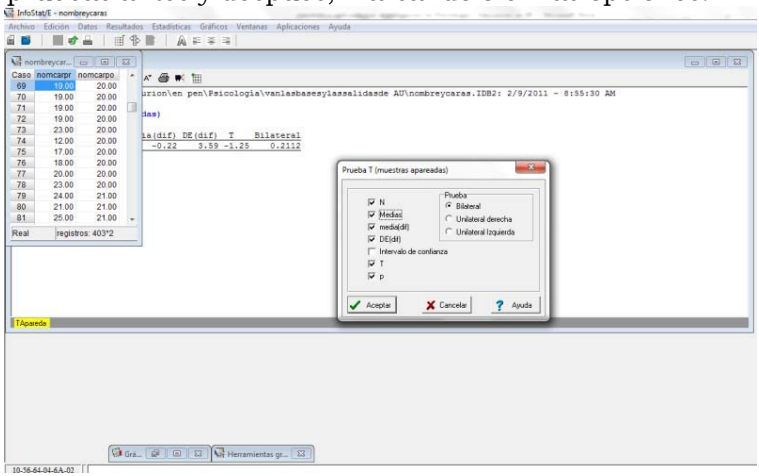


Por ahora solicitamos la prueba bilateral (H_1: $\overline{D} \neq 0$) y dejamos marcado lo que se ofrece por defecto. La salida es:

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	T	Bilateral
nomcarpr	nomcarpo	401	-0.22	3.59	-1.25	0.2112

Es posible solicitar a InfoStat® que también nos muestre las medias de las pruebas antes y después, marcándolo en las opciones:



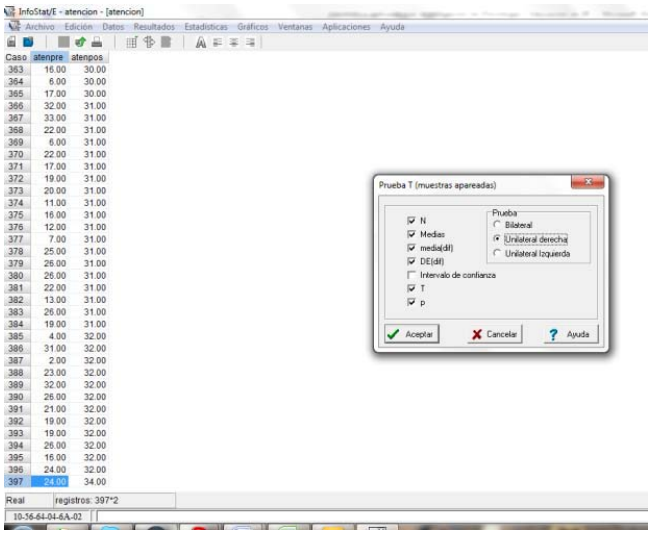
Ahora la salida resulta:

Obs(1)	Obs(2)	N	media(dif)	Media(1)	Media(2)	DE(dif)	T	Bilateral
nomcarpr	nomcarpo	401	-0.22	23.36	23.58	3.59	-1.25	0.2112

Vemos que las medias difieren en poca cantidad: de 23,36, el puntaje sube a 23,58. Esa apreciación intuitiva sobre la pequeñez de la diferencia es confirmada por el valor de probabilidad (bajo el rótulo Bilateral), que es alto (21%), muy superior a un nivel de significación del 5% o inclusive del 10%. Como resultado, no rechazaremos la H_0 y consideraremos que los puntajes no han cambiado. Decimos que el entrenamiento no ha tenido mayores efectos en esta habilidad tan compleja, como la de relacionar caras y nombres.

Ejemplo 11.8
Aplicada sobre la misma muestra de adultos mayores, es una prueba de atención que consiste en marcar en medio de una sopa de letras, sólo las letras zetas, en 30 segundos. Por cada acierto se le asigna un punto, sin descontar cuando cometen errores. Formularemos la prueba de modo unilateral, con la expectativa de hallar un aumento en el puntaje de atención; esperamos que el puntaje posterior sea mayor que el previo, por lo que la diferencia de la segunda medición

menos la primera dará positiva, luego $H_1: \overline{D} > 0$, es una prueba unilateral derecha. La lateralidad de la prueba se indica en la ventana de opciones:



Aplicada sobre 397 casos[91], la prueba da como resultado:

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	Media(1)	Media(2)	DE(dif)	T	p(Unilateral D)
atenpos	atenpre	397	5.19	23.04	17.85	7.38	14.00	<0.0001

Los nombres de las variables son atenpre (atención pretest, *Obs(2)*) y atenpos (atención postest *Obs(2)*)[92]. La primera lectura muestra que las medias son apreciablemente diferentes, el puntaje cambió de 17,85 a 23,04, es decir un aumento de 5,19 puntos, que juzgaríamos como elevado. En efecto, el valor de probabilidad de la última columna (*p(unilateral D)*) indica que la probabilidad de haber hallado esta diferencia solo por azar es menor a 1 en 10.000, lo que nos lleva a

[91] El número de casos difiere porque no siempre todas las personas completan íntegramente todas las pruebas, lo que da lugar a que haya casos perdidos, sea porque solo se cuenta con la medición previa o solo con la posterior.

[92] Observemos que la llamada *Obs(1)* es el resultado del postest y la *Obs(2)* es el puntaje del pretest, para esta elección hay que tener en cuenta la forma en que el programa hace la resta: siempre opera como *Obs(1)-Obs(2)*, debe elegirse cómo llamar a cada una para que la resta sea en el orden requerido. La elección se realiza en la primera ventana de la prueba, al momento de ingresar las variables, la primera que se ingresa es *Obs(1)*, la segunda *Obs(2)*.

rechazar la H_0 y concluir que los puntaje efectivamente se incrementan luego del entrenamiento.

Resumen de pruebas sobre dos muestras tratadas en el capítulo

Parámetro	Estimador	Estadístico de prueba	Supuestos
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{comb} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$; $gl = n_1 + n_2 - 2$ $s_{comb}^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}$	Grupos independientes Distribución normal en la dos poblaciones ó $n_1 > 30$ y $n_2 > 30$ Varianzas iguales
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$	Grupos independientes Distribución normal en la dos poblaciones ó $n_1 > 30$ y $n_2 > 30$ Varianzas diferentes
$P_1 - P_2$	$\hat{p}_1 - \hat{p}_2$	$z = \frac{(\hat{p}_1 - \hat{p}_1)}{\sqrt{\frac{\hat{p}_1 * (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 * (1 - \hat{p}_2)}{n_2}}}$	Grupos independientes y $n_1 > 100$ y $n_2 > 100$
$\overline{D}$	$\bar{d}$	$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}; gl = n - 2$	Distribución normal de d, ó $n > 30$

Actividad práctica de repaso 11

1. Los resultados que se muestran a continuación provienen de una muestra de alumnos que cursaron Psicoestadística en 2009. Se comparan los resultados del primer parcial entre los turnos mañana y tarde. Considere la siguiente salida:

Prueba T para muestras Independientes

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	T	gl	p-valor	prueba
Turno	Primero	{M}	{T}	115	228	6,63	6,60	0,15	341	0,8838	Bilateral

a. ¿Cuál es la unidad de análisis?
b. ¿Qué variable se analiza?
c. ¿Cuántos casos se consideran?
d. ¿Cuáles son los grupos que se comparan?
e. ¿Cuántos casos hay en cada grupo?
f. ¿Cuánto valen los promedios muestrales?
g. ¿Cuál es la hipótesis nula de la prueba?
h. ¿Cuál es la hipótesis alternativa?
i. ¿Cuántos son los grados de libertad de la prueba?
j. ¿Por qué?
k. ¿Cuál es la lectura del valor p?
l. Si el nivel de significación es del 5%, ¿cuál es la conclusión?
m. ¿En qué habría consistido cometer ETII en esta prueba?

2. Ahora nos interesa conocer si la proporción de aprobados del primer parcial difiere entre los turnos mañana y tarde. Para ello, recodificamos las notas asignando el valor cero (0) a los menores a cuatro y uno (1) a los cuatro y superiores. La variable dicotómica que así resulta se llama *aprobó primero* y se puede tratar con el mismo procedimiento que una cuantitativa. La salida InfoStat ® para esta prueba es:

Prueba T para muestras Independientes

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	T	gl	p-valor	prueba
Turno	aprobó primero	{M}	{T}	115	232	0,87	0,88	-0,38	345	0,7065	Bilateral

a. ¿Qué son las "medias", cuyos valores son 0,87 y 0,88?
b. ¿Cuál es la hipótesis nula de la prueba?
c. ¿Cuál es la hipótesis alternativa?
d. ¿Cuál es la lectura del valor p?
e. Si el nivel de significación es del 5%, ¿cuál es la conclusión?

3. A fin de analizar si las notas del segundo parcial disminuyeron significativamente respecto del primero, ahora comparamos el resultado del primer parcial con el del segundo alumno por alumno y obtenemos:

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media (dif)	DE(dif)	T	p(Unilateral D)
Primero	Segundo	343	1,66	2,28	13,46	<0,0001

a. ¿En qué se diferencia esta comparación de medias del caso 1?
b. ¿Cuál es la hipótesis nula de la prueba?
c. ¿Cuál es la hipótesis alternativa de la prueba?
d. ¿Cómo se llama el estimador puntual?
e. ¿Qué significa el valor p<0,0001?
f. En base a ese resultado, ¿cuál es la decisión, al 5%?
g. ¿Qué habría significado cometer ETI en esta prueba?

4. En una prueba estandarizada de atención, se comparan los puntajes de personas de dos grupos de edades: el primero compuesto por quienes tienen entre 50 y 65 años y el segundo por mayores de 65. Se cree que las personas de mayor edad podrían alcanzar puntajes más bajos en esta prueba.
a. ¿Cuál es la hipótesis nula de la prueba?
b. ¿Cuál es la hipótesis alternativa?
c. La lateralidad de la prueba es……
d. ¿En qué consistiría cometer ETI en esta prueba?
e. ¿En qué consistiría cometer ETII en esta prueba?

Capítulo 12: Comparación de más de dos grupos

Eduardo Bologna

Desde el comienzo hemos destacado la naturaleza multicausal de los fenómenos que son de nuestro interés y la imposibilidad de conocer todos los factores que inciden en la ocurrencia de lo que observamos. Iniciamos el tratamiento de este problema con los capítulos sobre relaciones entre variables, cuando señalamos que el camino de la explicación consiste en identificar variables que contribuyan de manera parcial a dar cuenta de las variaciones en otra variable. Denominamos antecedentes (o independientes) a las primeras y consecuente (o dependiente) a la segunda. En este capítulo volvemos sobre el problema de la multicausalidad, desde una perspectiva un poco diferente, pero sobre el final veremos que es un modo más completo de analizar relaciones entre variables.

Un concepto que será importante tener presente para este capítulo es el de coeficiente general de determinación. Recordemos que, para relaciones asimétricas, este coeficiente mide la proporción de los cambios de la variable dependiente que son explicados por la presencia de la independiente. En el contexto del análisis de regresión, cuando decimos "la proporción de los cambios" nos referimos a la parte de la varianza total que es explicada por el modelo. Para definirlo separamos la varianza total en una parte atribuible a la variable independiente a través del modelo y una parte "no explicada", que atribuimos a otros factores además del que se incluye en el modelo lineal. La técnica llamada Análisis de la Varianza (ANOVA o ANAVA o también one way) se basa en un procedimiento similar y tiene como objetivo el de determinar si varios grupos difieren en los valores promedio de una variable cuantitativa. Por ejemplo, si los pacientes sometidos a un tratamiento convencional difieren en el tiempo que tardan en ser dados de alta, de otros pacientes que reciben una terapia cognitiva y también de un tercer grupo de pacientes bajo tratamiento con psicofármacos. Si solo se tratara de dos grupos, podríamos usar la prueba t para diferencia de medias y una extensión de esa prueba consistiría en repetir la prueba t para los pares de grupos, por ejemplo; comparar tratamiento convencional con

terapia cognitiva, tratamiento convencional con psicofármacos y finalmente terapia cognitiva con psicofármacos. Además de engorroso (y peor si se trata de cuatro o más grupos), este proceder aumenta considerablemente el error de tipo I de la prueba en su conjunto.

La idea de base del procedimiento para comparar las medias de una variable cuantitativa entre varios grupos independientes (tres o más, ya que para dos grupos disponemos de la prueba *t*) es la de separar la variabilidad total de una variable (dependiente) en una parte que se atribuye a la pertenencia de los casos a los diferentes grupos y otra parte que se debe a otros factores. Esta última componente se considera como variabilidad individual, es la que no se explica por la pertenencia a los grupos.
El mejor modo de seguir los pasos del procedimiento será a través un ejemplo.

Ejemplo 12.1
Supongamos que se pretende conocer si los niños de primer grado aprenden a leer más rápidamente con un método de enseñanza, al que llamaremos A, que con otro, B. Tomando como variable al tiempo de aprendizaje, puede calcularse el promedio de tiempo necesario para adquirir ciertas destrezas básicas por parte de niños que aprendieron con cada uno de los diferentes métodos y luego comparar esos valores promedio entre los grupos, sometiendo los datos a una prueba *t* de diferencia de medias para muestras independientes. Esta prueba plantea, como hipótesis nula que las medias poblacionales son iguales, lo que en este caso particular implica que los diferentes métodos no dan lugar a diferencias en los tiempos de aprendizaje de la lectura. Así entonces, si el valor t que se obtiene del estadístico de prueba es mayor que el valor crítico establecido de antemano (de acuerdo al nivel de significación elegido), se rechazará la hipótesis nula y se concluirá que las diferencias son significativas, lo cual implica que con los diferentes métodos, los niños aprenden en tiempos diferentes. De lo contrario, si el valor obtenido de *t* resulta inferior en valor absoluto al de *t* crítico, se aceptará H_0, concluyéndose que no hay diferencias significativas entre los tiempos empleados para aprender a leer con los diferentes métodos. Hasta aquí, lo que sabemos de comparación de dos grupos.

El vocabulario del ANOVA

Ampliemos ahora el problema considerando que se trata de determinar si existe diferencia entre los tiempos empleados para

aprender a leer por niños a los que se enseña según tres métodos diferentes: A, B y C. Antes de comenzar con la descripción del análisis de la varianza, estableceremos cierto vocabulario inicial. La muestra está constituida por todos los elementos a los que se aplica la prueba, distribuidos entre los diferentes grupos bajo análisis. La denominación clásica (que tiene origen en el diseño experimental), es la de **tratamiento** para cada uno de los grupos en que se separó la muestra. En el ejemplo anterior, los tratamientos están constituidos por los tres métodos de enseñanza de lectura que se comparan. Considerando que el método constituye una variable (nominal), y que A, B y C son sus categorías, suele denominarse **factor** a esta variable independiente y **niveles del factor** a sus categorías.

Se llama **factor** a la variable independiente.
Los **niveles del factor** son las categorías de esa variable, también llamadas **tratamientos** o simplemente, **grupos**.

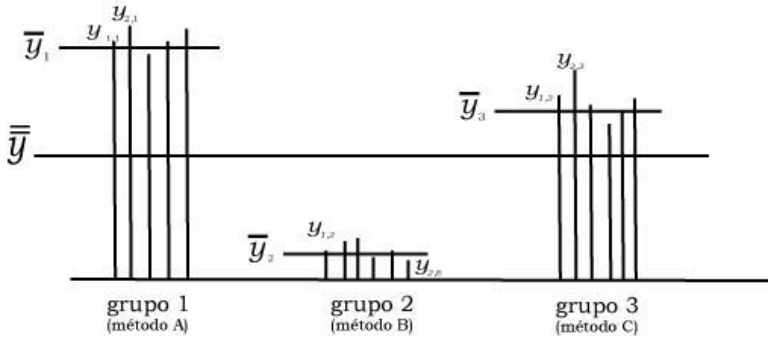
En nuestro ejemplo, el factor es el método de enseñanza y sus niveles cada uno de los métodos puestos a prueba.
Cada niño tarda su tiempo en aprender y las diferencias que se observen en esos tiempos estarán dadas parcialmente por los diferentes tratamientos a los que han sido sometidos (diferentes métodos de aprendizaje) y parcialmente por otros factores (características individuales por ejemplo); a estos últimos no los vamos a identificar en nuestro análisis. La idea del ANOVA es la de separar la varianza total de la muestra en dos componentes aditivas (que se suman): una de ellas debida a los efectos de la variable independiente y otra que mide los efectos debido al resto de las variables que no se consideran en el modelo.
La variable dependiente es aquella cuyos cambios se busca explicar y debe ser siempre cuantitativa, ya que sobre ella se calcularán medias y varianzas. Esta variable se conoce como **variable de salida**, porque se interpreta como el resultado de la independiente.
Mantendremos la notación que usamos en regresión, de llamar x a la variable independiente e y a la dependiente, por lo que x es el factor, e y la variable cuantitativa cuyos cambios se analizan. Los nombres para cada elemento son los siguientes:
$y_{i,j}$: es el valor que asume la variable dependiente (de salida). Se trata del tiempo que tardó en aprender uno de los niños. El primer subíndice i, representa al número de individuo, el segundo j, es el tratamiento (grupo) al que ese individuo pertenece. Así, el puntaje $y_{3,2}$ es el obtenido por el tercer individuo del segundo grupo. En este

ejemplo, es el tiempo que tardó en aprender el tercer niño que usó el método B (segundo grupo).
$\bar{y}_j$: representa el promedio del grupo *j*. Por ejemplo $\bar{y}_3$ es tiempo promedio requerido para aprender a leer por todos los niños del grupo al que se enseñó con el método C (tercer grupo).
$\bar{\bar{y}}$: es el promedio general de todos los individuos de todos los grupos. El tiempo promedio que tardaron en aprender todos los niños, sin importar con qué método hubiesen estudiado.
Continuaremos usando la letra n para indicar el total de individuos que participan en la muestra.
Con la letra k representaremos el número de grupos que tomamos (o el número de niveles del factor).
Para la cantidad de individuos en cada grupo también usamos n, pero con un subíndice que señale a qué grupo se refiere, así n_1 es la cantidad de casos en el primer grupo y de manera general, n_j será la cantidad de casos en el grupo *j*, por lo que será:

$$n = \sum_{i=1}^{k} n_i$$

Que indica que el total de casos es la suma de las cantidades que tiene cada grupo, desde el primero (*n=1*) hasta el último (*k-ésimo*)

El siguiente esquema ayuda a recordar los nombres de cada componente, para el caso de tres grupos:



Cada línea vertical representa el valor de la variable dependiente para cada individuo: el tiempo que tardó cada niño en aprender a leer. Están reunidas en tres grupos, el primero tiene 5 casos, y el segundo y tercero tienen 6 cada uno. Por lo que:

$$k = 3; n = 17, n_1 = 5;\ n_2 = 6;\ n_3 = 6$$

La primera línea vertical, indicada como $y_{1,1}$ es el tiempo que tardó en aprender a leer el primer niño del grupo que aprendió con el método A. De manera análoga, la del lado se llama $y_{2,1}$ se refiere al segundo niño del mismo grupo y así los demás individuos de los otros dos grupos. Las líneas horizontales representan las medias de cada grupo ($\bar{y}_1$; $\bar{y}_2$; $\bar{y}_3$) y la media general $\bar{\bar{y}}$.

La descomposición de la variabilidad de la variable de salida

Para un individuo dado, su diferencia respecto del promedio general dependerá del tratamiento al que fue sometido y de sus características individuales. Por ejemplo, el primer niño del grupo 1 tardó un tiempo en aprender a leer que indicamos como $y_{1,1}$. Para responder por qué ese niño en particular tardó ese tiempo en aprender a leer, diremos que, en parte eso se debe al método con el que aprendió, ya que todo su grupo tarda más que el promedio general (porque la media del grupo es mayor que la media general: $\bar{y}_1 > \bar{\bar{y}}$). Pero además, ese niño está por encima de la media de su grupo ($y_{1,1} > \bar{y}_1$). De manera abreviada podemos escribir esto, para el primer niño del grupo 1 como:

$$y_{1,1} - \bar{\bar{y}} = (\bar{y}_1 - \bar{\bar{y}}) + (y_{1,1} - \bar{y}_1)$$

$(y_{1,1} - \bar{\bar{y}})$ es la diferencia entre el tiempo que tardó el sujeto *1,1* y el tiempo promedio empleado por el conjunto completo (los tres grupos juntos).

$(\bar{y}_1 - \bar{\bar{y}})$ es la diferencia entre el tiempo promedio empleado por el grupo al que pertenece el sujeto (grupo 1), y el promedio que tardó el conjunto completo.

$(y_{1,1} - \bar{y}_1)$ es la diferencia entre el tiempo que tardó el sujeto *1,1* y el tiempo promedio del grupo al que pertenece.

En esta expresión hemos descompuesto la distancia a la que se encuentra el individuo *1,1* de la media general, en dos distancias:
La primera es la que hay entre su grupo y la media general, esa distancia toma en consideración el efecto del grupo, es decir, en qué medida el método pudo haber incidido en el tiempo que se tarda para aprender a leer.
La segunda diferencia es la distancia entre el individuo *1,1* y la media de su propio grupo, esta distancia mide efectos individuales, debidos a otros factores que no son el método usado.

Así entonces, para dar cuenta de la diferencia a la que un sujeto se encuentra del promedio general, consideramos el aporte de la pertenencia al grupo (haber estudiado con determinado método) y las características individuales del sujeto.

Para llevar esta descomposición a una forma general, hacemos:

$$y_{i,j} - \bar{\bar{y}} = (\bar{y}_j - \bar{\bar{y}}) + (y_{i,j} - \bar{y}_j)$$

El primer miembro es la diferencia entre un valor particular y el promedio general de todos los grupos, mide todo lo que se aparta un individuo de la media. En el segundo miembro, el primer término es una medida de lo que se aleja el grupo completo (representado por su media), del promedio general; esta diferencia es la que atribuimos al tratamiento que hemos aplicado. El segundo término es la distancia a la que se encuentra el individuo del promedio de su propio grupo; esta diferencia mide cuánto se distingue él de otros que pertenecen al mismo grupo, por lo que se explica por características propias del sujeto al que se considera, no por la pertenencia al grupo.

Para hacer extensiva esta operación a todos los individuos de la muestra, sumaremos estas diferencias, pero como se trata de desvíos en torno a la media, será necesario elevarlas al cuadrado para que la suma no sea igual a cero. Puede demostrarse que la expresión toma la forma:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{i,j} - \bar{\bar{y}})^2 = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{\bar{y}})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y}_j)^2$$

El primer miembro:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{i,j} - \bar{\bar{y}})^2$$

Se llama **suma de cuadrados total** (*SCT*). Es la suma de los cuadrados (para evitar que se anule) de las distancias que separan a cada individuo de la media general. Tiene doble signo de suma, porque deben sumarse todos los casos de cada grupo (*i*, que va hasta el último de cada grupo, n_j) y luego sobre todos los grupos (*j*, que va hasta *k*, el último de los grupos).

El primer término del segundo miembro:

$$\sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{\bar{y}}\right)^2$$

Se llama **suma de cuadrados explicada** (*SCEx*, o también **entre grupos**). Es la suma de los cuadrados de las distancias de cada grupo (representado por su media) a la media general. Incluye un factor (n_j) que tiene en cuenta los diferentes tamaños de los tratamientos, asignándoles un peso diferencial a cada uno de ellos, al multiplicar por n_j se da más importancia cuantitativa a los grupos que tienen más observaciones.

El segundo término:

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{i,j} - \bar{y}_j\right)^2$$

Se llama **suma de cuadrados residual** (*SCRes*, a veces denominada **dentro de los grupos** o también **suma de cuadrados del error**). Contiene la variabilidad individual: lo que cada sujeto se aleja del promedio del grupo al que pertenece. Tiene también doble signo de suma, por la misma razón que el primero: incluir a todos los individuos de todos los grupos.

La expresión puede escribirse de manera abreviada:

$$SCT = SCEx + SCRes$$

Los elementos de esa relación miden respectivamente: variaciones totales, variaciones que pueden atribuirse al modelo que se pone a prueba (o explicadas por el modelo) y variaciones debidas a diferencias individuales (residuales o no explicadas por el modelo).
Los nombres que reciben estas sumas de cuadrados difieren según la bibliografía; por ejemplo, el programa InfoStat® llama *Total* a la primera, *nombre de la variable independiente* a la explicada y *error* a la residual. En nuestro ejemplo, la *SCExp* se denominaría *Método empleado para enseñar a leer*.

Dijimos al comienzo que la idea del análisis de la varianza consiste en comparar las varianzas originadas en los tres tipos de variación que se han señalado (total, explicada y residual), pero las sumas de cuadrados que acabamos de definir no constituyen aun varianzas, son sólo los numeradores de las mismas. Para obtener las respectivas

varianzas será necesario dividir las sumas de cuadrados por sus respectivos grados de libertad.
Para hacerlo será necesario recordar la breve introducción que hicimos a ese concepto: los grados de libertad dependen del número de variables aleatorias en cada una de las sumas de cuadrados y del número de parámetros que deben estimarse para su cálculo.

-Para la suma de cuadrados total hay tantos valores de y (la variable aleatoria) como casos en total, es decir hay n valores y sólo se estima la media general μ a través de $\bar{\bar{y}}$, los grados de libertad serán entonces n (variables aleatorias) menos uno (por el único parámetro que debe estimarse): *n-1*.

-Para la suma de cuadrados residual tenemos n valores de y, y k medias de tratamiento (cada una de las $\bar{y}_j$ estima una media poblacional de grupo μ_j) que deben estimarse, entonces los grados de libertad correspondientes serán: *n-k*.

-Para la suma de cuadrados explicada vemos que hay k variables (las medias de los grupos) y sólo un parámetro que se estima (la media poblacional, μ), por lo que los grados de libertad resultan: *k-1*.

Las varianzas son ahora los cocientes de las sumas de cuadrados divididas por los grados de libertad. Estas varianzas también se conocen como "cuadrados medios".
La varianza total:

$$s^2 = \frac{SCT}{n-1} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{i,j} - \bar{\bar{y}}\right)^2}{n-1}$$

Esta es simplemente la varianza muestral: la dispersión del conjunto completo, sin considerar las pertenencias a los grupos.

La varianza explicada:

$$s^2_{expl} = \frac{SCEx}{k-1} = \frac{\sum_{j=1}^{k} n_j\left(\bar{y}_j - \bar{\bar{y}}\right)^2}{k-1}$$

Es el estimador de la varianza que se atribuye a la pertenencia a los diferentes grupos.

Y la varianza residual:

$$s_{res}^2 = \frac{SCRes}{n-k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{i,j} - \bar{y}_j\right)^2}{n-k}$$

Es el estimador de la varianza debida a factores ajenos a los grupos, por ejemplo diferencias individuales.

Estas dos últimas varianzas reciben diferentes nombres según la bibliografía.
La "varianza explicada" puede aparecer como "varianza entre grupos", o también como "cuadrado medio explicado" o "cuadrado medio entre grupos".
La "varianza residual" puede encontrarse indicada como "varianza dentro de los grupos" o "cuadrado medio dentro", "cuadrado medio residual" o "cuadrado medio del error".

Los supuestos del ANOVA
La prueba de análisis de la varianza se realiza cuando se comparan grupos que son independientes entre sí y este es el primero de los supuestos que hacen válida la prueba.
El segundo supuesto es la normalidad de la distribución de la variable de salida en la población de donde provienen las muestras. Este supuesto puede reemplazarse por tamaños de muestra suficientemente grandes que permitan la aproximación normal usando el Teorema Central del Límite. Cuando las muestras son pequeñas, debe verificarse la normalidad, en el próximo capítulo indicaremos el procedimiento que se usa para hacerlo.
El tercer supuesto es la igualdad de las varianzas poblacionales de los grupos que se comparan, se denomina homocedasticidad. Sin embargo, cuando este supuesto no se cumple, existen procedimientos de corrección similares al de Welch–Satterthwaite (que usamos en la *prueba t* cuando las varianzas son diferentes), que permiten hacer el análisis de la varianza y comparar las medias de varias poblaciones.

La prueba de hipótesis sobre las medias de grupos

El objetivo de este análisis es saber si los diferentes tratamientos producen efectos diferentes, queremos saber si las medias de varias poblaciones difieren entre sí o bien si las diferencias observadas sólo se deben a efectos del azar. La hipótesis nula de esta prueba será entonces que no hay diferencias entre las medias de los k tratamientos, mientras que la hipótesis alternativa planteará que al

menos una de las medias difiere, esto se escribe de la siguiente manera:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1: al\ menos\ una\ de\ las\ medias\ difiere$$

Entonces la hipótesis nula será rechazada cuando las diferencias debidas a los tratamientos sean sustancialmente más importantes que las debidas a otros factores. Las diferencias debidas a los tratamientos están medidas con la s^2_{expl} y las diferencias debidas a otras factores se resumen en s^2_{res}. Entonces debemos decidir si la primera de estas varianzas es significativamente mayor que la segunda. Para compararlas tomaremos su cociente:

$$\frac{s^2_{expl}}{s^2_{res}}$$

Según el programa que se utilice para hacer los cálculos también podrá encontrarse expresado como cociente de los cuadrados medios:

$$\frac{CM_{entre}}{CM_{dentro}}$$

No olvidemos que estamos trabajando con una muestra aleatoria, por lo que el resultado que hallemos de esta comparación estará en parte determinado por las reales diferencias que haya entre las varianzas y en parte por azar. La pregunta tiene la misma forma que ya hemos visto para otras pruebas de hipótesis: ¿qué tanto más grande debe ser la parte explicada que la residual para que estemos autorizados a rechazar H_0? Volveremos a dar la respuesta en términos probabilísticos, porque, si se cumplen los supuestos, el cociente de las varianzas se distribuye con la distribución que mencionamos en el capítulo 6: la *F de Fisher*[93]. El estadístico de prueba resulta entonces:

$$F_{k-1,n-k} = \frac{s^2_{expl}}{s^2_{res}}$$

Como el interés está dado por saber si el numerador supera al denominador (o, lo que es lo mismo, si la varianza explicada es significativamente mayor que la residual), entonces se rechazará la hipótesis nula si se obtiene un valor observado "grande" del cociente de los cuadrados medios; en consecuencia, se tratará de una prueba unilateral derecha, para la que la zona de rechazo de H_0 se sitúa en los valores ubicados a la derecha del valor crítico de *F*.

[93] Cuyos grados de libertad son los del numerador y denominador del cociente.

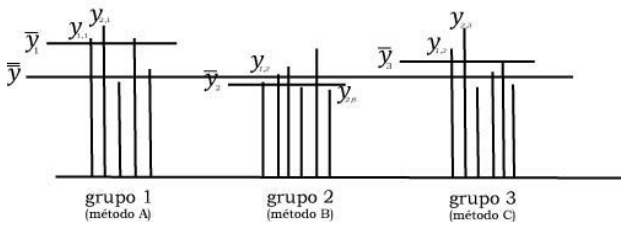
A fin de simplificar la presentación de los cálculos, se utiliza una tabla, conocida como tabla de ANOVA, como la siguiente:

Tabla de análisis de la varianza: disposición de los elementos que conducen a la comparación de la variabilidad entre grupos con la variabilidad dentro de los grupos.

Fuente de variación	Suma de cuadrados	grados de libertad	Varianzas	F observado	F crítico
Explicada	*SCExpl*	*k-1*	$s^2_{expl} = \frac{SCExpl}{k-1}$	$F_{obs} = \frac{s^2_{expl}}{s^2_{res}}$	$F_{k-1,n-k,1-\alpha}$
Residual	*SCRes*	*n-k*	$s^2_{res} = \frac{SCRes}{n-k}$		
Total	*SCT*	*n-1*			

Indicamos como valor crítico de *F* (con grados de libertad *k-1* y *n-k* en el numerador y denominador respectivamente), al correspondiente a un área acumulada de $1-\alpha$, dado que se trata de una prueba unilateral derecha. Así entonces el procedimiento consistirá en:
1. calcular las sumas de cuadrados
2. dividirlas por los correspondientes grados de libertad para obtener las varianzas
3. dividir estas varianzas entre ellas para obtener el valor de *F* observado, es F_{obs}
4. identificar el valor crítico de *F*, con los grados de libertad *k-1* en el numerador y *n-k* en el denominador, y un área acumulada de $1-\alpha$. Se trata de F_c
4. comparar F_{obs} con F_c. Si lo supera ($F_{obs} > F_c$), concluiremos que los tratamientos tienen efectos diferentes. De lo contrario, si el valor F_{obs} es inferior al puntaje crítico ($F_{obs} < F_c$), se acepta H_0 y la conclusión es que no hay diferencia entre los efectos producidos por los diferentes tratamientos.

Cuando no hay diferencia entre los grupos, o dicho de otro modo, cuando la diferencia apreciada entre las medias puede atribuirse al azar, entonces el esquema que muestra la posición relativa de los grupos respecto del promedio general, tiene forma:



En el que se ve que aunque las medias muestrales difieran, los individuos pueden estar por encima o por debajo de la media general, sin definir una tendencia que permita afirmar que los grupos difieren.

Ejemplo 12.2
Sea un grupo de 15 estudiantes a los que se divide en tres grupos. A fin de dejar claro que no es necesario que los grupos sean de igual tamaño, consideremos que los grupos contienen 4, 5 y 6 individuos cada uno. Los denominaremos grupo 1, 2 y 3. Los estudiantes de cada uno de los grupos han asistido a un curso de traducción de un idioma extranjero (que se supone conocían con anterioridad) durante un periodo de dos meses. Los cursos se dictaron con diferentes técnicas y se quiere evaluar si esas técnicas conducen a aprendizajes significativamente diferentes. La hipótesis nula para esta prueba afirmará que no hay diferencias entre los resultados obtenidos por los alumnos que aprendieron con las distintas técnicas; que es equivalente a decir que las eventuales diferencias observadas en los logros promedio de los tres grupos sólo se deben al azar. La hipótesis alternativa dice que los resultados difieren entre sí debido a los diferentes procedimientos utilizados. Expresado con la notación habitual:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1: al\ menos\ una\ de\ las\ medias\ difiere$$

La siguiente tabla muestra los puntajes obtenidos en una prueba de traducción estandarizada (se trata de un tipo de prueba que evalúa de manera equivalente a todos los individuos) aplicada a los quince alumnos que constituyen la muestra completa. Estos puntajes son los valores de la variable de salida. Vamos a suponer que en la población, la variable de salida tiene distribución normal.

Grupos (tratamientos)		
1	2	3
8	7	5
9	8	6
9	9	6
10	8	5
	7	7
		9
$n_1 = 4$	$n_2 = 5$	$n_3 = 6$

Usando la notación que presentamos, tenemos *n=15* (es 4+5+6), es el tamaño de la muestra. *k=3*, es el número de tratamientos. De modo

que los grados de libertad serán, para los desvíos entre tratamientos (o explicados) *k-1=3-1=2* y para los desvíos residuales (o dentro de los tratamientos) será *n-k=15-3=12*. Para el total es *n-1=14*, pero no lo usamos.

Calculamos la media para cada uno de los grupos. Para el primero de ellos resulta $\bar{y}_1 = \frac{8+9+9+10}{4} = 9$
De la misma manera se obtienen las medias de los otros grupos:

$$\bar{y}_2 = 7{,}80$$
$$\bar{y}_3 = 6{,}33$$

Tomando en consideración a la muestra completa, la media general

$$\bar{\bar{y}} = \frac{8+9+9+10+7+8+9+8+7+5+6+6+5+7+9}{15} = 7{,}53$$

A partir de estos resultados podremos calcular la contribución de cada fuente de variación a la varianza total.
La suma de cuadrados entre tratamientos (explicada) será:

$$SCExpl = \sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{\bar{y}}\right)^2 =$$
$$= 4 * (9 - 7{,}53)^2 + 5 * (7{,}8 - 7{,}53)^2 + 6 * (6{,}33 - 7{,}53)^2 = 17{,}6$$

Para la suma de cuadrados dentro de los tratamientos (residual) tendremos:

$$SCRes = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{i,j} - \bar{y}_j\right)^2 =$$
$$= (8-9)^2 + \cdots + (10-9)^2 + (7-7{,}8)^2 + \cdots + (7-7{,}8)^2 + (5-6{,}33)^2 + \cdots + (9-6{,}33)^2$$

Esta operación muestra con más claridad el sentido de la doble suma: primero se suma dentro de cada grupo y luego se suman los grupos.

La suma de cuadrados total será:

$$SCT = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{i,j} - \bar{\bar{y}}\right)^2 =$$
$$= (8-7{,}53)^2 + \cdots + (10-7{,}53)^2 + (7-7{,}53)^2 + \cdots + (7-7{,}53)^2 + (5-7{,}53)^2 + \cdots + (9-7{,}53)^2 = 33{,}73$$

Podemos verificar que se cumple que:

$$SCT = SCExpl + SCRes = 17{,}6 + 16{,}13 = 33{,}73$$

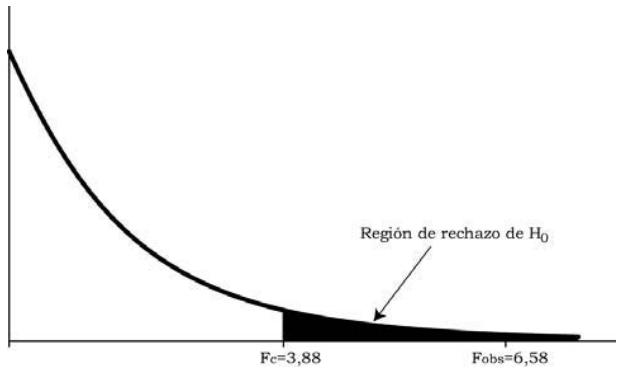
Ahora pueden calcularse, las varianzas entre tratamientos (explicada) y dentro de ellos (residual):

$$s^2_{expl} = \frac{17{,}65}{2} = 8{,}82$$
$$s^2_{expl} = \frac{16{,}13}{12} = 1{,}34$$

La tabla de análisis de la varianza toma ahora la siguiente forma:

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianzas	F observado	F crítico
Explicada	*SCEx=17,65*	*k-1=2*	$s^2_{expl} = 8,82$	$F_{obs} = \frac{8,82}{1,34} = 6,58$	$F_c = 3,88$
Residual	*SCRes=16,13*	*n-k=12*	$s^2_{res} = 1,34$		
Total	*SCT=33,73*	*n-1=14*			

Gráfico 1: Ubicación del punto crítico y del valor observado, en una distribución *F* con 2 grados de libertad en el numerador y 12 en el denominador, para α=0,05



El punto crítico es el valor más allá del cual se rechazará la hipótesis nula. Dado que el valor observado de F es 6,58, y supera al valor crítico, la decisión es la de rechazar Ho y concluir que, a un nivel de significación del 5% al menos uno de los tres métodos difiere de los demás. Dicho de otra manera, concluimos que las diferencias de puntaje obtenidas en la prueba no se deben solo a que los alumnos difieran en sus características individuales, sino que se explican por los diferentes métodos utilizados.

Los mismos datos analizados por InfoStat® ofrecen la siguiente salida:

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
puntaje	15	0,52	0,44	15,39

Cuadro de Análisis de la Varianza (SC tipo I)

F.V.	SC	gl	CM	F	p-valor
Modelo	17,60	2	8,80	6,55	0,0120
metodo	17,60	2	8,80	6,55	0,0120
Error	16,13	12	1,34		
Total	33,73	14			

La lectura de la salida es la siguiente
FV: abrevia "fuente de variación", son los elementos en que se descompone la varianza.
SC: sumas de cuadrados.
gl: grados de libertad.
CM: cuadrados medio, son las varianzas de cada componente.
F: el valor observado del estadístico de prueba.
p-valor: el valor de probabilidad asociado al F_{obs}. Como en las otras pruebas de hipótesis, indica la probabilidad de hallar un valor como el obtenido o más extremo, si la hipótesis nula fuera verdadera. Su pequeño valor (1,2%) indica que, de ser verdadera H_0 (todas las medias iguales) habría sido muy improbable encontrar un valor *F* de 6,55 ó mayor, por eso decidimos rechazar H_0 de igualdad de todas las medias, y concluimos que al menos una media difiere de las demás.

El cuadro de la salida mostrada tiene las dos primeras filas iguales, eso es porque solo hemos trabajado con una variable explicativa, solo nos interesa saber si el método hace las diferencias. En un diseño más complejo, podríamos haber introducido también el tipo de docente y observar los efectos de esta otra variable sobre los puntajes de la prueba. En ese caso la primera fila, que se llama *modelo* habría incluido los efectos conjuntos del método y el tipo de docente.

Antes del cuadro de ANOVA, la salida ofrece una descripción de los datos: tamaño de muestra (15 casos) y el cálculo del coeficiente general de determinación (R^2), al que conocimos en el análisis de regresión. Este coeficiente resulta de haber dividido la suma de cuadrados explicada por la suma de cuadrados total:

$$R^2 = \frac{SCExpl}{SCT} = \frac{17{,}60}{33{,}73} = 0{,}52$$

Veamos qué significado tiene: dado que la suma de cuadrados total está descompuesta en dos sumas de cuadrados, el cociente de una parte (la SC_{expl}) en el todo (SCT) indica la proporción que representa esa parte respecto del total. Así, el cociente es la proporción de la variabilidad total (medida por las sumas de cuadrados), que representa la parte explicada. En otros términos es la parte de la variabilidad en los resultados, que se explica por la variable bajo análisis: el método. Por eso podemos leer el coeficiente general de determinación como "de todos los factores que explican las diferencias de resultados que los alumnos obtienen en la prueba, el método aporta el 52%". O bien "el 52% de las diferencias en los puntajes que los alumnos alcanzan se explica por el método que usaron para aprender". De modo que el 48% restante de la variabilidad depende de otros factores.
Se trata de un coeficiente que aporta mucha información sobre la relación que existe —en nuestro ejemplo—, entre los resultados y el método. De manera análoga puede usarse esta prueba y el coeficiente R^2 para evaluar el aporte relativo que diferentes variables hacen a la explicación de otras.

Otras pruebas de hipótesis que apelan a la partición de la varianza

Mencionamos brevemente un conjunto de pruebas que usan la misma idea que hemos expuesto en este capítulo para analizar relaciones entre variables en diferentes casos.

Anova de dos factores (o dos vías)
Este procedimiento evalúa el efecto conjunto de dos variables nominales sobre una cuantitativa. Se evalúa la parte de la varianza total que es explicada por cada una de las dos variables y por la interacción de ambas.
En una investigación (Ridao García y Gil Flores, 2002) se compara el rendimiento de alumnos (evaluado como el promedio de calificaciones) que van a colegios con jornada continua (solo clases por las mañanas) y con jornada partida (que agregan clases algunas tardes)[94]. La

[94] Para la definición de esta modalidades los investigadores indican: "Los centros que formarán parte de la muestra habrían de representar a cada una de las dos modalidades organizativas consideradas en la población: jornada escolar de cinco mañanas (jornada continua) y jornada escolar de cinco mañanas y un número de entre dos y cuatro tardes (jornada partida)" (Ridao García y Gil Flores, 2002, p. 146)

comparación solo requiere de una prueba *t de student*, porque son solo dos grupos. Pero los investigadores suponen que podría haber también diferencias según se trate de establecimientos públicos o privados, lo cual introduce una segunda variable explicativa de los rendimientos. Para este problema se requiere un ANOVA de dos factores, correspondiente a las dos variables independientes que explicarían el rendimiento.

Análisis de Covarianza (ANCOVA)
Agrega al análisis de la varianza (de uno o más factores) el análisis del efecto de una o más variables continuas. Se supone un efecto lineal de la variable continua sobre la variable dependiente que es constante para los diferentes niveles de los factores. Es decir que la misma función lineal que relaciona la variable independiente continua con la variable de salida es válida para todos los niveles de las otras variables.
Suele decirse que este análisis "limpia" de los efectos de variables independientes continuas para mejorar la calidad del análisis de los efectos de los factores. En el ejemplo anterior, los autores podrían haber decidido poner a prueba el efecto de la edad (variable cuantitativa) también como factor explicativo, en ese caso, el ANCOVA es adecuado.

Análisis multivariado de la varianza (MANOVA)
Usa el mismo razonamiento que el análisis de la varianza pero la variable de salida es un vector, es decir que tiene varias componentes. Esto puede resultar, por ejemplo, de un test que arroje varias puntuaciones para cada sujeto. Ese conjunto de puntuaciones es el que se desea comparar entre varios grupos.
Ampliando el ejemplo anterior, los investigadores habrían podido evaluar el rendimiento de los alumnos con más de un valor, no solo el promedio de calificaciones. Por ejemplo, cada alumno podría ser puntuado según: su promedio de calificaciones, el número de materias que debió rendir al año anterior y las inasistencias en el año. Se dispondría así del rendimiento medido a través de tres variables más simples, y se lo trata como un vector de tres componentes. Para comparar el *rendimiento* así construido según la *jornada* (continua o partida) y la *gestión* (pública o privada) de la escuela, se usa un MANOVA.

Actividad práctica de repaso 12

Se comparan las notas de un examen entre alumnos que cursaron la materia en los turnos mañana, tarde y noche.

1. Indique cuáles son:
a. La variable dependiente, o de salida.
b. El factor.
c. Los niveles del factor.
d. La hipótesis nula.
e. La hipótesis alternativa.

2. La siguiente es la salida (incompleta) de InfoStat®

Cuadro de Análisis de la Varianza

F.V.	SC	gl	CM	F	p-valor
Modelo			33,65		0,0003
TURNO			33,65		0,0003
Error	909,11	224			
Total	976,41	226			

a. Complete los valores faltantes en la tabla de ANOVA
b. ¿Cuál es la decisión sobre H_0?
c. Redacte una lectura del valor *p*

3. Calcule e interprete el coeficiente general de determinación.

Capítulo 13: Pruebas sobre asociación entre variables

Eduardo Bologna

En los capítulos 4 y 5 analizamos relaciones entre variables con datos de muestras: las describimos, observando en qué medida se asocian las variables relevadas. De acuerdo al nivel de medición, calculamos diferentes coeficientes para medir el grado (o intensidad) de la asociación. En este capítulo, y luego de haber visto los procedimientos de inferencia, nos interesará generalizar las conclusiones obtenidas en estos análisis a toda la población de referencia. Disponemos ahora de las herramientas necesarias para hacer esa generalización. Para los coeficientes *r de Pearson* y r_s *de Spearman*, veremos que la lógica para la prueba de hipótesis es la que ya conocemos, solo cambiará la forma de calcular el estadístico de prueba. Para niveles de medición más bajos (ordinales o nominales), haremos una introducción a las pruebas no paramétricas.

Medida de la asociación en variables cuantitativas

Empezaremos por el último de los coeficientes que tratamos en el capítulo 5: el coeficiente de correlación lineal *r de Pearson*. Recordemos que se trata de un número comprendido entre -1 y 1 cuyo signo indica si se trata de una relación directa o inversa, y cuyo valor absoluto mide la intensidad de la asociación; cuanto más cercano es a 1 ó a -1, tanto más intensa es la asociación; por el contrario, si es cercano a 0 (cero) la asociación es tenue o muy débil. Los valores -1 y 1 —que no son observables en la realidad—, constituyen los extremos máximos del coeficiente y serían indicativos de una asociación lineal "perfecta". El valor 0 indica la ausencia de asociación y tampoco es posible que aparezca cuando se analizan datos reales.

Recordemos que el coeficiente de Pearson se calcula usando las transformaciones a puntaje *z* de los *n* valores de las dos variables cuya relación se analiza, luego esos puntajes se multiplican para obtener *r* con la siguiente expresión:

$$r = \frac{\sum_{i=1}^{n} z_{x_i} * z_{y_i}}{n-1}$$

Si ha sido obtenido en una muestra representativa, el coeficiente de correlación de Pearson allí calculado es un estimador de un coeficiente de correlación poblacional, es decir que mide la intensidad de la asociación entre las variables en la población; así como $\bar{x}$ es el estimador de μ y $\hat{p}$ el estimador de *P*. Manteniendo al notación anterior, usaremos la letra griega ρ (rho) para referirnos al coeficiente de correlación paramétrico (o poblacional).
Razonando como lo hicimos con la media en el capítulo sobre distribuciones en el muestreo: si se extraen todas las muestras posibles de una población dada y en cada una se calcula *r*, la media de todos los coeficientes de correlación encontrados, coincide con el coeficiente de correlación paramétrico. A esa media, que no calculamos, porque no es posible extraer "todas las muestras de la población", la llamamos esperanza.

Dicho de otra manera, r es un estimador insesgado de ρ: $E(r) = \rho$.
Por su parte, la varianza de r es:

$$\sigma_r^2 = \frac{1-r^2}{n-2}$$

Por lo que el error estándar del estimador tiene la forma:

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}}$$

Y, tal como sucedió con la media y la proporción, el Teorema Central del Límite indica que a medida que aumenta el tamaño de las muestras, la distribución de los r tiende a ser normal. Si podemos suponer que en la población, la variable bajo análisis tiene distribución normal, estamos autorizados para usar la distribución t cuando la muestra es pequeña. Cuando aumenta el tamaño de la muestra esta distribución tiende a la normal. En consecuencia, y como sucedió con la media, usaremos distribución *t de Stdent* siempre.
Para esta prueba de hipótesis, la distribución t tiene grados de libertad que se calculan como el tamaño de la muestra menos dos:

$$gl = n - 2$$

Conociendo la distribución de probabilidades para r podemos construir el estadístico de prueba restando el estimador menos el parámetro y dividiendo por el error estándar del estimador:

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

La hipótesis de carácter conservador (aquella que señala la no-diferencia) será la que afirme que en la población no hay relación entre las variables, por lo que si en la muestra hallamos un valor de r no nulo, habrá sido por azar. Nuevamente se trata de decidir si un valor muestral observado es evidencia suficiente para rechazar la H_0; a la que escribiremos:

$$H_0: \rho = 0$$

La hipótesis alternativa puede ser bilateral:

$$H_1: \rho \neq 0$$

o unilateral derecha:

$$H_1: \rho > 0$$

o unilateral izquierda:

$$H_1: \rho < 0$$

Y en todos los casos transformaremos el coeficiente r observado a puntaje t, haciendo:

$$t = \frac{r_{obs} - \rho}{\sqrt{\frac{1 - r_{obs}^2}{n - 2}}} = \frac{r_{obs}}{\sqrt{\frac{1 - r_{obs}^2}{n - 2}}}$$

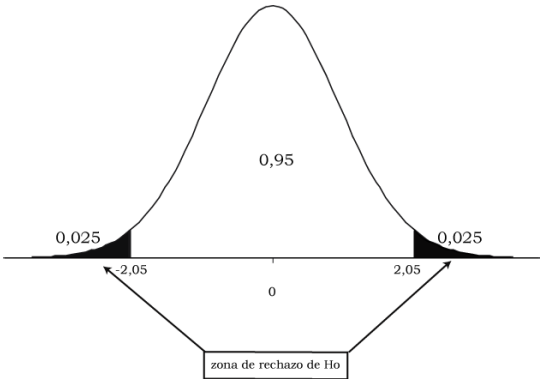
La última igualdad se debe a que la hipótesis nula afirma que $\rho = 0$.

Ejemplo 13.1
Se observa la relación entre el tiempo que los alumnos dedicaron a preparar el examen de una materia y la calificación obtenida en ese examen, a partir de una muestra de 27 estudiantes. Se obtiene un coeficiente de correlación lineal de Pearson de $r_{obs} = +0{,}37$. El signo positivo de este coeficiente indica que la relación es directa, por lo que, en la muestra, los alumnos que dedican más tiempo a la preparación, tienden a tener notas más altas. El valor absoluto indica que se trata de una relación moderada entre las dos variables. Ahora nos interesa probar si este resultado es suficiente evidencia para creer que en la población, las dos variables están relacionadas, es decir, que en la población el coeficiente de correlación no es cero. La

pregunta es si hemos encontrado 0,37 porque en la población las variables están efectivamente correlacionadas o solo por las variaciones propias del procedimiento aleatorio de muestreo, es decir, solo por azar.
Las hipótesis correspondientes a este problema son:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Es bilateral porque nuestro interés es saber si ρ difiere de cero.
Fijamos el nivel de significación en el 5% y hallamos en una distribución *t de Student* con 25 grados de libertad (*n-2=27-2=25*) los valores críticos correspondientes son ±2,05, gráficamente:



El área sombreada constituye la probabilidad extrema de 0,05, repartida en dos colas de 0,025 cada una. La zona de rechazo de H_0 se constituye por el conjunto de los valores *t* mayores a 2,05 así como los menores a -2,05.

Ahora transformamos el valor de *r* obtenido a puntaje *t*, calculando el estadístico de prueba:

$$t = \frac{r_{obs}}{\sqrt{\frac{1 - r_{obs}^2}{n-2}}} = \frac{0{,}37}{\sqrt{\frac{1 - 0{,}37^2}{27-2}}} = 1{,}99$$

Un valor que no se encuentra en la zona de rechazo, por lo que la decisión es la de aceptar H_0 y concluir que no hay suficiente evidencia para afirmar que haya una correlación lineal no nula entre el tiempo que los alumnos dedican a preparar una materia y la nota que obtienen en el examen. Diremos que el coeficiente de correlación

hallado en la muestra no difiere significativamente de cero ó, más simplemente, que no es significativo.

Como siempre puede hacerse en las pruebas de hipótesis, vamos a llegar a la misma conclusión calculando el *valor p* asociado al coeficiente *r* muestral. Se trata de encontrar la probabilidad de hallar un valor de *r* como el observado o uno más extremo que él, si la hipótesis nula fuera cierta. Como la prueba es bilateral debemos tratar a la expresión "más extremo" como incluyendo dos posibilidades, que *r* sea mayor que el observado o menor que su opuesto, por lo que la probabilidad que buscamos se escribe:

$$P(r < -r_{obs} \cup r > r_{obs} / H_0 \; es \; verdadera) = P(r < -0{,}37 \cup r > 0{,}37 / \rho = 0)$$

que se lee como "la probabilidad que *r* sea menor que el opuesto al valor observado o mayor que el valor observado, dado que la hipótesis nula es verdadera"
Usando el estadístico de prueba, transformamos los dos valores de r_{obs} en puntajes *t* y entonces esa probabilidad, expresada en términos de *t*, nos queda:

$$P(t < -1{,}99 \cup t > 1{,}99)$$

en la que no incluimos la condición que H_0 sea verdadera ($\rho = 0$) porque está implícita en la distribución *t* de *r*. Cuando la solicitamos a InfoStat®, esa probabilidad nos da 0,0568. Debido a que es mayor a nuestro nivel de significación ($\alpha = 0{,}05$), la decisión es la de no rechazar H_0. La lectura de este valor *p* es "si en la población no existiera correlación entre las variables, la probabilidad de haber hallado un valor como el observado o más extremo que él sería de 0,0569. Consideramos a este valor como elevado, por lo que no rechazamos la hipótesis de ausencia de correlación".

Dos consecuencias de este resultado:
Primera: un coeficiente que, a nivel muestral habríamos juzgado como moderado no puede generalizarse como significativo a toda la población. Esto se debe principalmente al reducido tamaño de la muestra. Así, para que este valor de *r* hubiese representado una asociación significativa en la población, habría sido necesario que proviniese de una muestra de mayor tamaño.
Segunda: el valor de t_{obs} es cercano al punto crítico (1,99 frente a 2,05), de modo que es poco lo que le falta para que la hipótesis sea rechazada. En el procedimiento a través del valor *p*, eso equivale a que dicho valor de probabilidad apenas supera al nivel de significación

establecido: 5,6% frente a 5%. Aun así, nuestro criterio fijado a priori indica que no debemos rechazar H_0.

Solicitada a InfoStat, esta operación da el siguiente resultado:

Coeficientes de correlación

Correlación de Pearson: Coeficientes\probabilidades

	horas	promedio
horas	1,00000	0,05608
promedio	0,37195	1,00000

La expresión *Coeficientes\probabilidades* indica que en la diagonal secundaria de la tabla debe leerse: primero el coeficiente, en la intersección de las dos variables (promedio y horas) y luego la significación (el valor *p*) en la otra intersección de las variables (horas y promedio). La lectura es entonces que la correlación muestral entre las dos variables es 0,37195 y que la probabilidad de que ese resultado haya sido encontrado por puro azar es 0,05608. A un nivel de significación del 5% no se rechaza la H_0 según la cual las variables no están correlacionadas. Aunque sí se rechaza si se fija un nivel de significación "más tolerante" del 10%.

Correlación entre variables ordinales

El coeficiente *r de Pearson* solo puede interpretarse si proviene de variables medidas a nivel intervalar o proporcional, es decir, métricas. Cuando trabajamos con variables ordinales, disponemos de otra medida de la asociación, el coeficiente de correlación por rangos r_s *de Spearman*. Recordemos que proviene de transformar en rangos los n valores de las dos variables ordinales cuya relación se analiza. La resta, caso por caso, de esos rangos, da lugar a las diferencias, llamadas *d*, con las que se calcula el coeficiente:

$$r_s = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Los valores se interpretan como los de *r de Pearson*.

La prueba de hipótesis que permite generalizar su valor a una población de referencia, casi no difiere de la que acabamos de describir para *r*. Llamaremos ρ_s al coeficiente de Spearman paramétrico y r_s al muestral. El error estándar de este estimador tiene la misma forma que el de Pearson:

$$\sigma_{r_s} = \sqrt{\frac{1 - r_s^2}{n - 2}}$$

y cuando la muestra tiene al menos 10 observaciones, se distribuye con una distribución t con n-2 grados de libertad[95]. En consecuencia, el estadístico de prueba será:

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}}$$

Ejemplo 13.2
Sea que se interroga a 20 alumnos de primer año de una carrera universitaria sobre las razones de su elección de carrera y a partir de las respuestas se construye un índice que clasifica el interés en 1. Muy fuerte, 2. Fuerte, 3. Débil, 4. Muy débil. Los resultados de este índice se ponen en correspondencia con el orden de mérito alcanzado en el ingreso a la carrera universitaria (1. Primero, 2. Segundo, etc.). En el análisis de la relación entre estas dos variables ordinales obtenemos un coeficiente de correlación de Spearman de 0,82, que indica, en la muestra, una relación positiva e intensa entre el interés y el orden de mérito en el ingreso. Preguntamos si este resultado, obtenido sobre 20 casos, nos autoriza a afirmar que existe una asociación entre las dos variables más allá de la muestra observada. Formularemos las hipótesis correspondientes a una prueba bilateral, porque queremos probar si el coeficiente es significativamente diferente de cero.

$$H_0: \rho_s = 0$$
$$H_1: \rho_s \neq 0$$

Fijamos el nivel de significación en el 5% y hallamos en una distribución t con 18 grados de libertad (n-2=20-2=18) los valores críticos correspondientes son ±2,10. Calculamos el estadístico de prueba:

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}} = \frac{0{,}82}{\sqrt{\frac{1 - 0{,}82^2}{20 - 2}}} = \frac{0{,}82}{0{,}135} = 6{,}08$$

[95] Kendall (1948) citado por Siegel (1956)

Este valor se ubica en la región de rechazo de H_0, con lo que la decisión es la de rechazar H_0 y concluir que hay evidencia para creer que la correlación a nivel poblacional no es nula.

Cuando se pide a InfoStat®, la salida tiene el mismo formato que la del coeficiente de Pearson:

Coeficientes de correlación
Correlación de Spearman: Coeficientes\probabilidades

	interés	orden de mérito
interés	1,00000	0,00036
orden de mérito	0,81842	1,00000

Nuevamente, la barra oblicua entre coeficiente y probabilidades indica que 0,81842 es el coeficiente r_s (que habíamos hallado en 0,82 antes) y 0,00036 es el valor de probabilidad asociado. El pequeño valor de éste último es señal de lo escasamente probable que resulta que este resultado provenga del azar, es decir, es evidencia para creer que en la población, la asociación efectivamente existe.

Pruebas no paramétricas

Los requisitos que hemos solicitado hasta este punto para los procedimientos vistos, son exigentes. Por ejemplo, en las *pruebas t* debemos suponer que la variable tiene distribución normal en la población. El Teorema Central del Límite nos dice que si las muestras son lo suficientemente grandes, la distribución muestral tiende a ser normal, por lo que puede eliminarse el supuesto de normalidad en la población si las muestras tienen suficiente tamaño, aunque esto implica mayores costos, que no siempre es posible afrontar. Todas las pruebas que hemos visto hasta este momento hacen supuestos acerca de la población, que son condiciones que debe cumplir la distribución de las variables bajo análisis en la población, estos supuestos a veces se cumplen y otras no. Los resultados que se obtengan de esas pruebas dependen del cumplimiento de esas condiciones. En algunos casos es posible poner a prueba la veracidad de esas exigencias, para ver si se cumplen, en otros casos, solo es posible "suponer" que es así. Por esa razón los resultados pueden ser aproximados o directamente incorrectos, si hay violaciones graves de sus condiciones de aplicación.

Además, los cálculos hechos en las pruebas mencionadas, requieren un nivel de medición alto, para poder calcular medias y varianzas. Hemos resuelto parcialmente el problema cuando, al tratar con variables nominales, usamos la proporción de casos en alguna categoría. Sin embargo no hemos resuelto aún el problema de analizar relaciones entre variables cualitativas, a las que no puede calcularse la media ni la varianza, pero que a menudo aparecen en nuestros análisis.
Para este tipo de problemas, que son muy frecuentes tanto en Psicología como en Educación (pocos casos y variables que no son métricas) existe un conjunto de pruebas llamadas **pruebas no paramétricas**. Por oposición a ellas, todas las pruebas presentadas hasta aquí son **pruebas paramétricas**, que quiere decir que especifican ciertas condiciones que deben cumplir los parámetros de la población de la que se extrae la muestra.

Son **pruebas no paramétricas** las pruebas de hipótesis que no especifican condiciones sobre los parámetros de la población de la que proviene la muestra.

La limitación de las pruebas no paramétricas respecto de las paramétricas, es que tienen, a igual nivel de significación e igual tamaño de muestra, menor potencia. Eso significa que, para obtener la misma potencia en una prueba no paramétrica que en una paramétrica, es necesario usar más casos.
Si el problema es el del nivel de medición de las variables —que en la mayoría de los casos no es métrico—, puede resolverse apelando a una prueba no paramétrica y lograr resultados de la misma calidad, aunque a un mayor costo por la mayor cantidad de casos necesarios.
Si el problema es el tamaño de la muestra, es decir, si tenemos muy pocos casos observados y no puede suponerse distribución normal en la población, entonces no hay alternativa y debe usarse indudablemente una prueba no paramétrica.
En este capítulo solo desarrollaremos tres pruebas no paramétricas basadas en el puntaje chi cuadrado, puede consultarse el manual de Siegel (1956) para una presentación muy completa, aunque no actualizada a procedimientos informáticos.

Las pruebas *ji cuadrado* (o *chi cuadrado*)
Hemos ya presentado el puntaje *ji cuadrado* en el capítulo 5, allí fue usado para derivar medidas de la asociación entre dos variables nominales (V de Cramer, C de Pearson). El término *ji cuadrado* volvió

a aparecer en el capítulo 6, como un modelo especial de probabilidades. Ahora relacionaremos esos dos usos del puntaje *chi cuadrado*, en primer lugar para analizar la eventual independencia entre dos variables (prueba de independencia de atributos), luego para evaluar si una distribución se ajusta a un modelo predicho (prueba de bondad de ajuste) y finalmente para comparar la tendencia central entre variables de nivel ordinal (prueba de la mediana).

Prueba de independencia de atributos

En el capítulo 5 presentamos el puntaje *ji cuadrado* (χ^2) como medida de la distancia que hay entre una distribución de frecuencias bivariada observada y la correspondiente distribución de frecuencias esperadas. Estas frecuencias esperadas son las que se esperarían observar si las variables fueran independientes. Así, cuanto más alejadas se encuentren las frecuencias observadas de las esperadas, tanto más grande será el puntaje χ^2 y más alejadas de la independencia estarán las variables que se analizan. Dijimos también que estar alejadas de la independencia implica que existe alguna asociación entre ellas. El valor mínimo de este puntaje es cero, que solo podría alcanzarse si todas las frecuencias observadas coincidieran con las esperadas y estaríamos en el caso de una independencia perfecta; una situación muy improbable de hallar en la realidad. El puntaje χ^2 carece de un valor máximo, puede ser indefinidamente grande, dependiendo no solo de lo alejadas que están las frecuencias observadas de las esperadas, sino también de la dimensión de la tabla y de la cantidad de casos.

Dado que las frecuencias observadas que están en la tabla bivariada provienen de una muestra, necesitamos poder generalizar el resultado que hallemos a la población de referencia. Es decir que ahora, en el contexto de la inferencia, nos preguntamos ¿Cómo debería ser de grande el puntaje χ^2 encontrado para que consideremos que las variables se alejan lo suficiente de la independencia? Por la redacción de la pregunta puede verse que tratamos con una prueba de hipótesis, nos estamos preguntando ¿A partir de qué valor podemos considerar que χ^2 es significativo? La respuesta dependerá del valor que asuma χ^2 (que está influido por el tamaño de la muestra) y de la dimensión de la tabla. Para formular las hipótesis de esta prueba, recordemos que la H_0 es aquella que indica no-diferencia, el no-cambio, es la hipótesis "conservadora". Por el contrario, la hipótesis alternativa presenta una diferencia en algún sentido. En este problema, como vamos a tratar acerca de relaciones entre variables, la

H_0 indicará que no hay relación o, lo que es lo mismo, que las variables son independientes.

Ejemplo 13.3
Se trata analizar la posible relación entre el resultado del primer parcial de Psicoestadística (aprobado – no aprobado) y el turno en que los alumnos lo hicieron. Se construye una muestra con 180 alumnos seleccionados entre quienes hicieron el primer parcial en los últimos cinco años. Sea la siguiente la tabla de distribución de frecuencias observadas.

Tabla 1: Frecuencias conjuntas observadas del resultado del parcial y el turno en que se realizó:

		Turno			
		M	T	N	Total
Resultado del parcial	Aprobado	60	30	40	130
	No aprobado	30	10	10	50
	Total	90	40	50	180

Sobre esta tabla podemos –como hicimos en el capítulo 4—, calcular frecuencias relativas por columnas, dado que nos interesa saber si los resultados difieren según el turno. Pero ahora nos concentraremos en decidir si estas variables son independientes o no.
Para este análisis, ya sabemos calcular las frecuencias esperadas[96], que dan:

Tabla 2: Frecuencias conjuntas esperadas del resultado del parcial y el turno en que se realizó.

		Turno			
		M	T	N	Total
Resultado del parcial	Aprobado	65	29	36	130
	No aprobado	25	11	14	50
	Total	90	40	50	180

Las frecuencias esperadas nos permiten calcular el puntaje[97] χ^2, que en este caso es $\chi^2 = 3{,}05$. Este es el número que habíamos usado para calcular los coeficientes de asociación (*V de Cramer* y *C de Pearson*).

[96] Que se calculan como $f_{ij} = \frac{f_i * f_j}{n}$

[97] Haciendo: $\chi^2 = \sum \frac{(f_o - f_o)^2}{f_e}$

Ahora, nuestro interés es el de generalizar a toda la población, se trata de alcanzar una conclusión acerca de la independencia o no de las dos variables, no restringida a estos 180 casos, sino general. No debemos olvidar que los datos disponibles son muestrales y, por el modo en que se seleccionan los casos de la muestra, dependen del azar.

Nos preguntamos: ¿el valor hallado para la medida sintética de la distancia a la que se encuentran las tablas 1 y 2, que es el puntaje $\chi^2 = 3{,}05$, puede explicarse por azar, o representa un distanciamiento suficientemente grande como para atribuirlo a una relación entre las variables en la población?

Queremos entonces decidir si lo que hemos observado para estos alumnos (la muestra de quienes cursaron en los últimos cinco años, por ejemplo) es general, es decir, si vale para alumnos a los que no hemos observado. Por eso, la hipótesis nula de la prueba será:

> H_0: El resultado del parcial es independiente del turno en que se realiza[98]

que equivale a afirmar que no hay relación entre el turno y el resultado.

Su contrapartida, la H_1 dirá que:

> H_1: Existe relación entre el turno en el que se realiza el parcial y el resultado que se obtiene.

La pregunta será entonces si la evidencia hallada a partir de nuestros datos es suficiente para rechazar Ho y concluir que las variables están relacionadas o si, por el contrario, deberemos seguir sosteniendo que las variables no están relacionadas.

Conocemos del capítulo 6 que la variable aleatoria χ^2 tiene una distribución de probabilidad asimétrica y que su forma depende de los grados de libertad. Estos últimos, para tablas de doble entrada dependen del número de filas y de columnas que tenga la tabla, según:

$$gl = (f - 1) * (c - 1)$$

Donde gl son los grados de libertad, f es el número de filas de la tabla y c el número de columnas. Entonces conocemos el puntaje χ^2 y su distribución de probabilidad, con lo que podemos decidir si se trata de un valor extremo (muy poco probable si H_0 fuera cierta) o bien de un valor esperable. En el primer caso rechazaremos H_0 y concluiremos

[98] De mismo modo que sucede con las pruebas paramétricas, las hipótesis se refieren a toda la población, aunque aquí no aparezca el nombre de un parámetro reconocible, como μ o P.

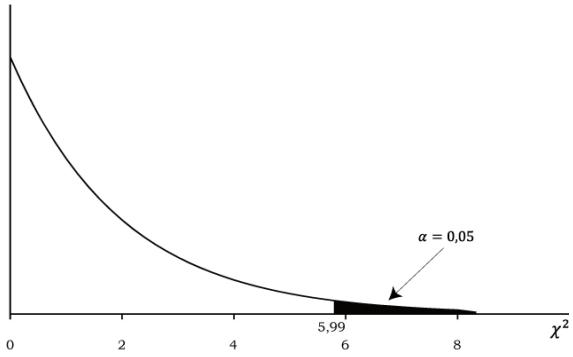
que hay relación entre el turno en que se hace el parcial y el resultado que se obtiene. En el segundo caso, aceptaremos H_0 y diremos que no hay evidencia para descartar la independencia, o que no hay pruebas para sostener que las variables estén relacionadas. Rechazaremos H_0 si el valor χ^2 encontrado es grande, porque eso es sinónimo de un gran alejamiento de nuestros datos respecto de la independencia. En razón de ello, ésta será siempre una prueba unilateral derecha: se rechaza para valores que excedan cierto límite.

El procedimiento para hacer la prueba puede ser el tradicional, que consiste en fijar de antemano el nivel de significación (α), luego determinar el (único) punto crítico y comparar si el valor observado de χ^2 excede ese punto crítico; si es así se rechaza H_0. O también puede usarse el *valor p* asociado al puntaje χ^2 observado, y si esa probabilidad es menor al nivel de significación, se rechaza H_0.
En nuestro ejemplo, los grados de libertad son:

$$gl = (f-1) * (c-1) = (2-1) * (3-1) = 1 * 2 = 2$$

Fijamos un nivel de significación $\alpha = 0{,}05$ y, usando una hoja de cálculo, hallamos el punto crítico que corresponde para esa área (superior) y esos grados de libertad bajo la curva de la distribución χ^2, que resulta ser 5,99, éste es el puntaje crítico de χ^2, por lo que lo llamamos χ^2_c. La representación gráfica de la zona de rechazo de Ho es entonces:

Gráfico 1: Ubicación del punto crítico que deja un 5% de área superior en una distribución χ^2 con 2 grados de libertad.



De acuerdo a este gráfico, la región de rechazo de H_0 es la que se encuentra a la derecha de 5,99, formada por todos los valores de χ^2 que superen a 5,99. Nuestros datos ofrecieron un puntaje $\chi^2_{obs} = 3{,}05$, que no pertenece a la zona de rechazo de H_0, por lo que nuestra decisión es la de no rechazar H_0 y concluir que no hay relación entre las variables; o bien que no hay evidencia suficiente para descartar la independencia entre el resultado del parcial y el turno en que éste se realiza.

Si preferimos usar el valor de probabilidad (*valor p*) para tomar la decisión y dar a nuestros resultados mayor claridad, entonces haremos como en las pruebas paramétricas: informando la probabilidad de hallar un valor de χ^2 como el observado o más extremo que él. En esta prueba (de independencia de atributos), "más extremo" siempre quiere decir "mayor que", porque es una prueba unilateral derecha. Buscamos entonces la probabilidad que tiene la variable χ^2 con dos grados de libertad, de asumir un valor igual o mayor al que hemos observado, es decir:

$$P(\chi^2 \geq \chi^2_{obs}) = P(\chi^2 \geq 3{,}05)$$

Solicitada a InfoStat®, hallamos que esta probabilidad vale 0,2176. Este es el llamado *valor p* o *valor de probabilidad*. Lo leemos diciendo que, si las dos variables fueran independientes, la probabilidad de hallar un puntaje χ^2 como el observado o más extremo que él es de 0,2176. Dado que es una probabilidad alta (sustancialmente más grande que 0,05 que suele usarse como criterio de rechazo), consideramos a éste como un resultado altamente probable, de ser cierto que las variables son independientes.

Sobre estos mismos datos, la salida InfoStat® muestra:

Frecuencias absolutas
En columnas: turno

resultado	mañana	tarde	noche	Total
aprobado	60	40	30	130
no aprobado	30	10	10	50
Total	90	50	40	180

Frecuencias esperadas bajo independencia
En columnas: turno

resultado	mañana	tarde	noche	Total
aprobado	65,00	36,11	28,89	130,00
no aprobado	25,00	13,89	11,11	50,00
Total	90,00	50,00	40,00	180,00

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	3,05	2	0,2180
Chi Cuadrado MV-G2	3,10	2	0,2119
Coef.Conting.Cramer	0,09		
Coef.Conting.Pearson	0,13		

Disponemos aquí de las frecuencias observadas (la tabla de contingencia con nuestros datos) y la tabla de frecuencias esperadas si fueran independientes. Luego aparece el puntaje χ^2 (3,05) y su valor de probabilidad asociado (*p=0,2180*). No haremos lectura del resto de la información que provee la salida.

Pruebas de bondad de ajuste

El cálculo del puntaje χ^2 ofrece la posibilidad de hacer comparaciones entre frecuencias observadas (reales, provenientes de la recolección de datos) y frecuencias esperadas bajo diferentes supuestos. Hasta el momento las esperadas lo han sido bajo la hipótesis de independencia, ya que calculamos las frecuencias que debería haber en cada celda si las variables fueran independientes. Pero la condición bajo la cual se esperan determinadas frecuencias puede ser otra y el puntaje χ^2 también permite medir esas distancias.

Ejemplo 13.4

La distribución de la condición que alcanzan los alumnos que cursan Psicoestadística ha sido, históricamente, 35% de promocionados, 35% de regulares y 30% de libres.

Consideremos el subconjunto de alumnos que estudian otra carrera además de Psicología. Una muestra de 142 de estos alumnos se distribuye, según condición, del siguiente modo:

Tabla 3: Distribución de frecuencias de los alumnos de Psicología que también estudian otra carrera, según condición:

condición	alumnos
Promocionados	53
Regulares	51
Libres	38
Total	142

La tabla anterior muestra las frecuencias observadas. Nos preguntamos si esta distribución se aleja significativamente de la tendencia general, o bien si está dentro de lo esperado. Esta pregunta puede reformularse en dirección a saber si los datos observados se "ajustan" a la distribución general o se apartan de ella. Expresado en términos de hipótesis, la H_0 afirmará que no hay diferencia:

> H_0: La distribución de promocionados, regulares y libres de los alumnos que estudian otra carrera se ajusta a la distribución del total de alumnos.

Mientras que la H_1, afirmará lo contrario:

> H_1: La distribución de promocionados, regulares y libres de los alumnos que estudian otra carrera se aparta de la distribución del total de alumnos.

Si la distribución se mantuviera igual (si H_0 fuera verdadera), esperaríamos que los 142 alumnos que estudian otra carrera se distribuyeran en las tres categorías según estas proporciones: 0,35 - 0,35 - 0,30. La frecuencia que esperaríamos encontrar en la categoría promocionados es *0,35*142= 49,7* y del mismo modo con las demás categorías. Por lo que esperaríamos la siguiente distribución de frecuencias:

Tabla 4: Frecuencias esperadas bajo la hipótesis de ajuste a la distribución general:

condición	Alumnos
Promocionados	49,7
Regulares	49,7
Libres	42,6

Disponemos ahora de las dos tablas, una de frecuencias observadas y otra de esperadas y queremos evaluar si son similares o muy diferentes. El problema es el mismo que en la prueba de independencia de atributos, solo que las tablas son univariadas. Para medir la distancia entre las dos tablas disponemos del puntaje χ^2, que compara una a una las frecuencias de las celdas. Aplicada a las tablas de arriba, resulta:

$$\chi^2 = \sum \frac{(f_o - f_o)^2}{f_e} = \frac{(53 - 49{,}7)^2}{49{,}7} + \frac{(51 - 49{,}7)^2}{49{,}7} + \frac{(38 - 42{,}6)^2}{42{,}6} = 0{,}75$$

Del mismo modo que en la prueba de independencia de atributos, buscaremos un valor crítico de χ^2, para tomar la decisión. En este caso, los grados de libertad de la distribución χ^2 dependen del número de categorías de la tabla univariada, simplemente es el número de categorías menos uno:

$$gl = k - 1$$

En este ejemplo, las categorías son tres, por lo que *gl=2*. Con 2 grados de libertad y 5% de nivel de significación, el punto crítico es $\chi^2 = 5{,}99$, por lo que la región de rechazo es el conjunto de valores que superan 5,99. El valor observado se encuentra fuera de la región de rechazo, por lo que corresponde no rechazar H_0 y concluir que los alumnos que estudian otra carrera además de Psicología no muestran, en Psicoestadística, una distribución entre promocionados, regulares y libres que difiera significativamente de la tendencia general de quienes cursan esa materia.

Prueba de la mediana

Cuando es necesario comparar la tendencia central de dos distribuciones, se dispone de la *prueba t de diferencia de medias*. Sin embargo esa prueba no es válida cuando se trabaja con variables que tienen nivel ordinal, ya que allí no tiene interpretación la media ni tampoco la varianza. Si se deben comprar dos muestras en una variable medida a nivel ordinal se puede plantear la hipótesis que afirma que las medianas son iguales, como una equivalencia a la que, en variables métricas plantea que las medias de dos distribuciones son iguales:

$$H_0: M_{dn_1} - M_{dn_2} = 0$$

Frente a una hipótesis alternativa que usualmente es bilateral:

$$H_1: M_{dn_1} - M_{dn_2} \neq 0$$

El procedimiento consiste en calcular, en primer lugar, la mediana del grupo compuesto por los casos de las dos muestras, todos reunidos en una sola distribución, se la denomina "mediana combinada" y se indica M_{dn_c}. Luego se cuentan los casos de cada grupo que quedan por encima y por debajo de esa mediana combinada. Si la mediana de los dos grupos fuera la misma, se esperaría que aproximadamente la mitad de los casos de cada grupo queden por encima de la mediana combinada y la otra mitad por debajo. En la medida que los casos de

los grupos se aparten de esa forma de distribuirse habrá evidencia para creer que las medianas difieren.
La disposición de los datos se realiza en una tabla de dos por dos como la siguiente:

	Grupo 1	Grupo2	Total
Por encima de M_{dn_c}			
Por debajo de M_{dn_c}			
Total			

Bajo la hipótesis nula, esperaríamos que la cantidad de casos por encima y por debajo de la M_{dn_c} fuera la misma para los dos grupos. Dicho de otro modo, para aceptar la H_0 deberíamos hallar independencia entre la pertenencia a los grupos y la ubicación de los casos por encima y por debajo de M_{dn_c}. De este modo vemos que el problema puede tratarse como una prueba de independencia: si el puntaje χ^2_{obs} es tal que debe rechazarse la hipótesis de independencia, será —en esta prueba—, equivalente a rechazar la igualdad de las medianas.
Un problema que suele aparecer cuando se hace el recuento de casos por encima y por debajo de la M_{dn_c} es que algunos coincidan con ella, que no estén ni encima ni debajo. Si la cantidad total de casos con que se cuenta es grande, se pueden dejar de lado esos casos y solo contar los que efectivamente difieran (en más o en menos) de la M_{dn_c}. De lo contrario, puede hacerse el recuento de los casos que están por encima de la M_{dn_c} y los que no lo están, es decir que uno de los grupos cuenta la cantidad de casos que hay por encima de la M_{dn_c} y el otro los que la igualan o están por debajo de ella.

Ejemplo 13.5
Se desean comparar las calificaciones “de concepto” asignadas por docentes de escuelas primarias de dos grupos de alumnos. Dichas calificaciones, presentan categorías: 1. Excelente, 2. Muy bueno, 3. Bueno, 4. Satisfactorio, 5. No satisfactorio, por lo que son de nivel ordinal y no es posible realizar una prueba de diferencia de medias, en su reemplazo recurrimos a la prueba de la mediana. Se seleccionan 40 alumnos de cada docente y se relevan las notas de concepto de cada uno. Calculamos la mediana de los 80 alumnos juntos y obtenemos 3. A continuación contamos cuántos alumnos de la primera docente están por encima y por debajo de 3 y lo mismo para la segunda docente. La distribución queda del siguiente modo:

	Docente		
	1	2	Total
Por encima de M_{dn_c}	28	8	36
Por debajo de M_{dn_c}	8	24	32
Total	36	32	68

Aunque originalmente se relevaron 40 alumnos de cada docente, quedaron 36 de la primera y 32 de la segunda, porque cuatro casos de un grupo y ocho del otro coincidieron con la M_{dn_c} y fueron descartados, con lo que el total se redujo a 68 observaciones.
La sola inspección de la tabla sugiere que debe descartarse la igualdad de las medianas de los grupos, ya que hay concentración de casos en las celdas de la diagonal, que es una indicación de la relación que hay entre filas y columnas. En efecto, el puntaje $\chi^2_{obs} =$ 18,94 tiene, con 1 grado de libertad, una probabilidad asociada (*valor p*) de 0,0000135, que nos conduce a rechazar la hipótesis de igualdad de medianas.
Cuando esta operación se solicita a InfoStat®, la salida no presenta la tabla de clasificación de casos según su ubicación respecto de la M_{dn_c}, solo nos ofrece lo siguiente:

Prueba de la mediana para dos muestras

Clasific	Variable	n(1)	n(2)	Med	P(X1>Med)	P(X2>Med)	p(2 colas)
docente	concepto	40	40	3,00	0,70	0,20	<0,0001

que indica:
Clasific: el nombre de la variable que separa los grupos, en este caso, docente.
Variable: la característica que se compara.
n(1) y n(2): la cantidad de casos en cada uno de los grupos.
Med: la M_{dn_c}
P(X1>Med) y *P(X2>Med)*: la proporción de casos de cada grupo que superan a la M_{dn_c}, en este caso el 70% del primer grupo y el 20% del segundo.
P(2 colas): el valor p para prueba bilateral, obtenido a través de la prueba χ^2. Esta probabilidad no se presenta de manera exacta sino, en este caso, solo señalando que es menor que una diezmilésima, suficiente evidencia para rechazar H_0 y concluir que la diferencia de las medianas es significativa.

Apéndice

Cálculo de Pruebas Estadísticas utilizando Software Especializado: Aplicaciones con InfoStat

Leonardo Medrano

Introducción

A lo largo del presente libro se han desarrollado diferentes contenidos vinculados a la explicación de las bases lógicas y matemáticas de diferentes procedimientos estadísticos. Sin embargo, para lograr un uso efectivo de las aplicaciones estadísticas se requiere de conocimientos básicos para el manejo efectivo de software especializado en estadística (Kazdin, 2001). En la actualidad resulta difícil imaginar a un investigador del comportamiento humano realizando análisis estadísticos sin la ayuda de un software especializado. En efecto el uso de este tipo de programas tiene considerables ventajas con respecto al cálculo manual, ya que permite reducir el tiempo dedicado al análisis cuantitativo, aumentar su precisión, editar información, realizar representaciones gráficas y obtener salidas para elaborar informes, entre otras funciones (Manzano, Varela, García & Pérez, 1999).

Cabe destacar que el conocimiento sobre el uso y manejo de software estadístico simplemente complementa el conocimiento adquirido sobre las bases lógicas de cada procedimiento, vale decir, no lo reemplaza. Difícilmente puede utilizarse correctamente un programa estadístico si desconocemos las bases lógicas de cada procedimiento y simplemente nos limitamos a "clickear" esperando obtener algún resultado interesante. Como señala Gardner (2003), la expansión de los programas estadísticos ha incrementado el mal uso de técnicas estadísticas. Lamentablemente se observa con mucha frecuencia a usuarios desprevenidos empleando procedimientos inadecuados para el problema que tratan de examinar o interpretando de manera errónea los resultados obtenidos por el programa. En este sentido puede decirse que los programas estadísticos son "buenos esclavos pero malos amos", sólo podremos hacer un uso adecuado de los mismos si conocemos las bases conceptuales de los procedimientos estadísticos utilizados.

El objetivo del presente apéndice es el de introducir al estudiante sobre el uso y manejo responsable de los programas estadísticos, para

lo cual se requerirá del constante repaso de las secciones anteriores del libro relacionadas con la pruebas estadísticas ejemplificadas. En primer lugar se presentarán algunas generalidades sobre los softwares estadísticos centrándonos sobre el programa InfoStat, y posteriormente se expondrán los pasos necesarios para realizar análisis estadísticos con este programa. Concretamente nos centraremos sobre tres pruebas paramétricas muy utilizadas en psicología: 1) el coeficiente de correlación de Pearson, 2) la prueba *t* de Student y 3) el análisis de varianza de una vía (ANOVA).

¿Qué es y cómo obtener un programa estadístico?

En términos generales un software o programa estadístico se refiere a un conjunto de programas de ordenador que dispone de herramientas para analizar, editar, modificar y gestionar datos (Manzano et al., 1999). Existe un gran número de paquetes estadísticos de calidad muy variable, probablemente el más utilizado en el ámbito de las ciencias sociales sea el SPSS (iniciales de Statistical Package for the Social Science) actualmente comercializado con el nombre de PASW. No obstante existe una gran cantidad de alternativas comerciales, como el InfoStat y el Statistica por ejemplo, y no comerciales, como el OpenStat y el Vista que permiten realizar complejos procesos de gestión, análisis y presentación de resultados estadísticos (Ledesma, 2006). Aunque cada programa tiene características propias, los procedimientos generales son muy similares. En el presente apéndice se trabajará con el programa InfoStat ya que se trata de un paquete sumamente completo y desarrollado de manera integral en nuestro medio.

Un problema habitual en la utilización de software estadístico es el conocimiento y la obtención de los mismos (Manzano y Tobio, 2003), más aun en el caso de programas gratuitas ya que al no ser comerciales su publicidad es escasa. Para solucionar este problema, se sugiere la visita a la siguiente página: http://statpages.org/javasta2.html#General. Esta página web contiene links para acceder software estadístico de diversas características. Desde esta dirección se puede obtener e instalar programas de estadística, bioestadística, epidemiología, psicometría y programas de asesoramiento metodológico, entre otros. En algunos casos el acceso es totalmente libre y gratuito, en otros se trata de demostraciones de tiempo limitado. Quizás una desventaja de este sitio web es que la búsqueda de programas se torna un tanto compleja debido a la gran cantidad de programas disponibles.

Para obtener el programa InfoStat de una manera rápida y efectiva se sugiere la visita a la dirección: www.infostat.com.ar. Desde esta página podremos bajar una versión estudiantil gratuita (figura 1):

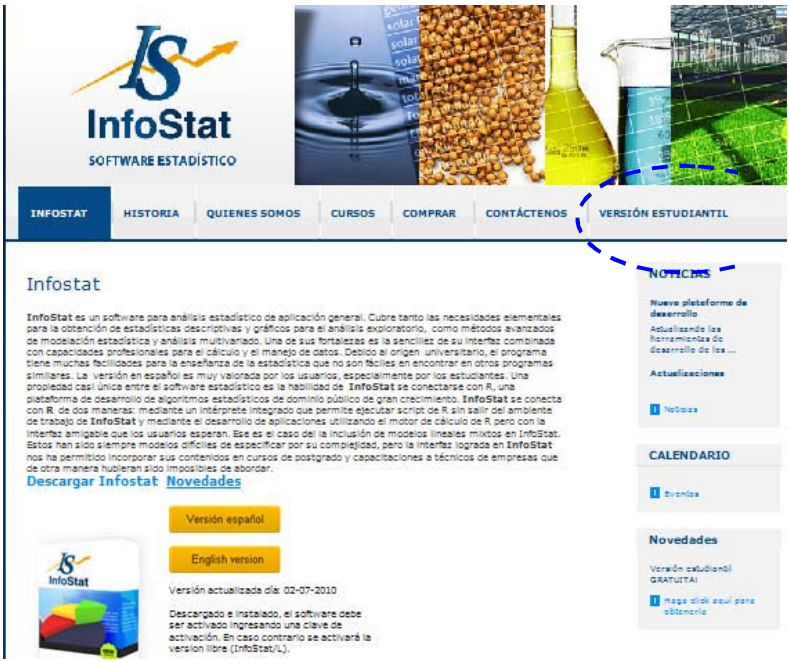


Figura 1: Página de inicio del sitio web oficial de InfoStat.

Una vez instalado el programa e iniciado el programa se podrá observar que los elementos que componen la apariencia inicial del editor de datos del InfoStat es similar a una hoja de cálculo. Tal como puede observarse en la figura 2, la ventana está compuesta por:

- *Barra de título,* con el nombre del fichero y los botones para minimizar, restaurar y cerrar la ventana.
- *Barra de herramientas* con sus respectivos menús de datos, tales como, *archivo, edición, datos, resultados, estadísticas,* entre otros.
- *Una barra de estado,* la cual suministra información sobre la actividad del programa, por ejemplo, los comandos que se estén ejecutando, los casos seleccionados o las variables trabajadas.

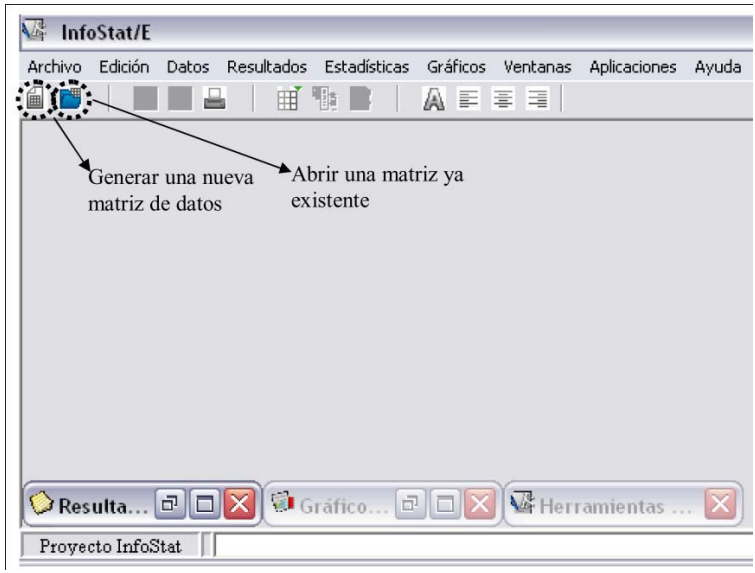
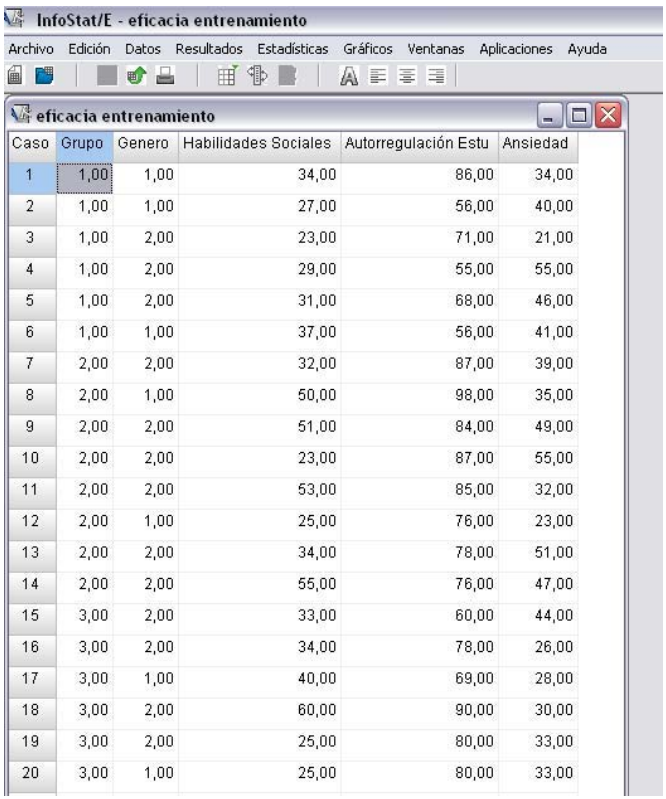


Figura 2: Apariencia Inicial del InfoStat.

Una vez que ingresamos al editor de datos podemos empezar a utilizar el programa de dos maneras, introduciendo datos en la matriz para generar una nueva base de datos, o bien, abrir una base de datos ya existente (figura 2). Para mostrar ejemplos de algunas aplicaciones que se pueden desarrollar con InfoStat, utilizaremos una base de datos simplificada de un estudio desarrollado por Medrano, Mirantes y Marchetti (2009) el cual consistió en evaluar la efectividad de un programa de intervención para ingresantes universitarios sobre sus habilidades sociales, autorregulación del estudio y ansiedad ante los exámenes (figura 3). La base o matriz de datos en este caso está compuesta por 20 filas (que representan la cantidad de casos de la muestra) y 5 columnas que representan las variables en estudio. Las variables consideradas son:

- *Grupo*: 1 = Grupo que recibió tratamiento; 2 = Grupo que no recibió tratamiento; 3 = Grupo placebo (piensa que recibió un tratamiento pero en realidad no lo recibió)
- *Género*: 1 = varón; 2 = mujer
- *Habilidades Sociales*: variable continua que refiere a la capacidad para realizar comportamientos socialmente competentes o exitosos
- *Autorregulación del Estudio*: variable continua que refiere a la capacidad para regular de manera autónoma el proceso de aprendizaje.

- *Ansiedad*: variable continua que refiere a los niveles de ansiedad experimentados frente a un examen.



Caso	Grupo	Genero	Habilidades Sociales	Autorregulación Estu	Ansiedad
1	1,00	1,00	34,00	86,00	34,00
2	1,00	1,00	27,00	56,00	40,00
3	1,00	2,00	23,00	71,00	21,00
4	1,00	2,00	29,00	55,00	55,00
5	1,00	2,00	31,00	68,00	46,00
6	1,00	1,00	37,00	56,00	41,00
7	2,00	2,00	32,00	87,00	39,00
8	2,00	1,00	50,00	98,00	35,00
9	2,00	2,00	51,00	84,00	49,00
10	2,00	2,00	23,00	87,00	55,00
11	2,00	2,00	53,00	85,00	32,00
12	2,00	1,00	25,00	76,00	23,00
13	2,00	2,00	34,00	78,00	51,00
14	2,00	2,00	55,00	76,00	47,00
15	3,00	2,00	33,00	60,00	44,00
16	3,00	2,00	34,00	78,00	26,00
17	3,00	1,00	40,00	69,00	28,00
18	3,00	2,00	60,00	90,00	30,00
19	3,00	2,00	25,00	80,00	33,00
20	3,00	1,00	25,00	80,00	33,00

Figura 3: Ejemplo de una matriz de datos del InfoStat.

Correlación de Pearson

Para evaluar la relación existente entre dos variables una prueba paramétrica habitualmente utilizada es el coeficiente de correlación de Pearson. Este coeficiente permite conocer la magnitud de la relación existente entre dos variables continuas y la dirección de dicha relación, la cual puede ser directa o inversa (ver el capítulo correspondiente a este coeficiente en el capítulo de "relaciones entre variables" del tomo II). Para calcular el coeficiente de correlación con InfoStat deben realizarse los siguientes pasos:

- Paso 1: Colocar el curso en *Estadísticas*, moverlo hacia abajo y colocarlo sobre *Análisis de correlación* y al desplegarse el menú clickear sobre *Coeficientes de correlación* (figura 4).

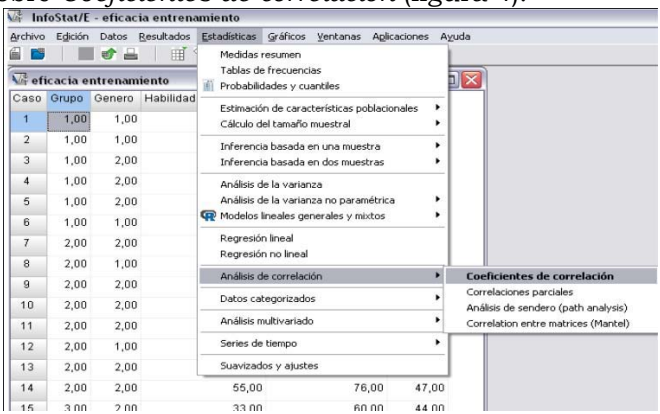


Figura 4: Calculo del Coeficiente de Correlación de Pearson con InfoStat.

-Paso 2: Se abrirá un menú que contiene las variables de la base de datos (figura 5). Se seleccionan las variables que se quieren correlacionar (en este caso se seleccionaron las variables *Habilidades Sociales* y *Autorregulación del Estudio*) y luego se selecciona el coeficiente de correlación correspondiente (en este caso el coeficiente de correlación de Pearson).

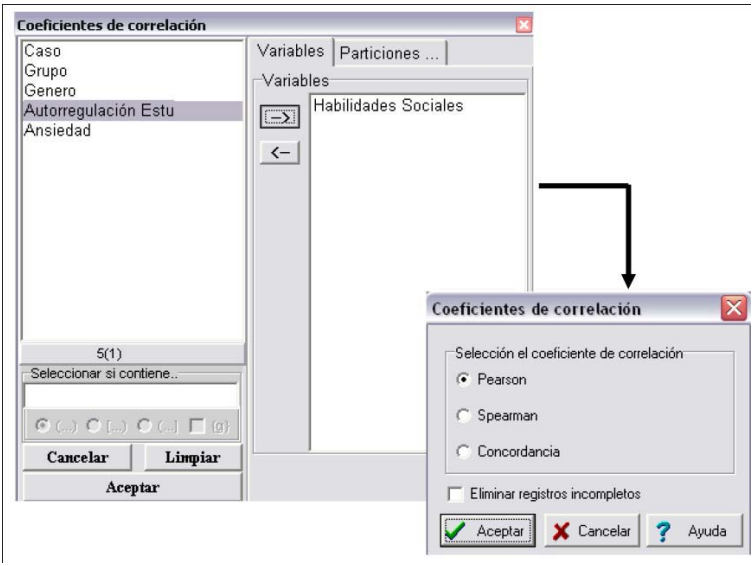


Figura 4: Calculo del Coeficiente de Correlación de Pearson con InfoStat

-Paso 3: Al clickear en *Aceptar* se desplegarán los resultados del análisis. Para realizar una correcta interpretación de los resultados se sugiere la lectura del capítulo correspondiente a "*relaciones entre variables*" del tomo II del presente libro.

Prueba *t* de Student para muestras Independientes

Este procedimiento estadístico es muy utilizado en psicología cuando se quiere determinar si existen diferencias entre dos grupos independientes. Esta prueba paramétrica permite comparar por ejemplo si personas que han recibido un tratamiento para dejar de fumar consumen menos cigarrillos que personas fumadoras que no han hecho el tratamiento, o bien si los pacientes de una clínica poseen más síntomas depresivos que personas no hospitalizadas. En este caso se realizará una comparación entre hombres y mujeres considerando los niveles de ansiedad que experimentan frente a un examen.

-Paso 1: Colocar el curso en *Estadísticas*, moverlo hacia abajo y colocarlo sobre *Inferencias basadas en dos muestras* y al desplegarse el menú clickear sobre *Prueba t* tal como se muestra en la figura 5.

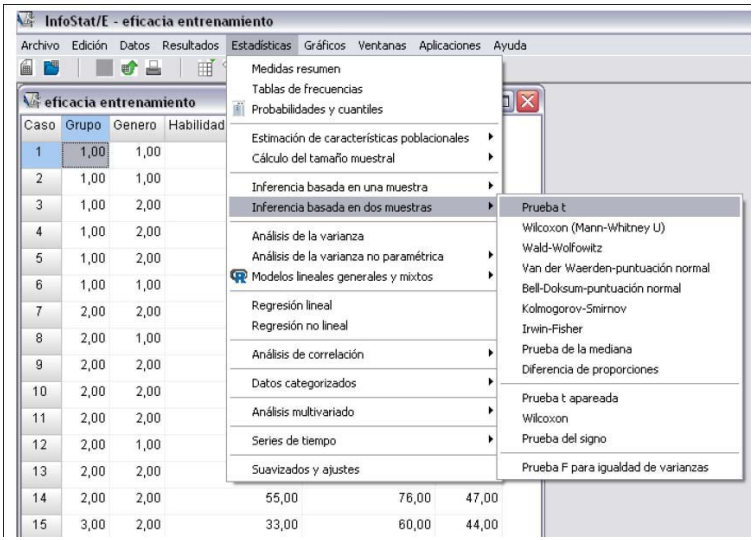


Figura 5: Calculo de Prueba t para muestras independientes con InfoStat

-Paso 2: Se abrirá un menú que contiene las variables de la base de datos (figura 6). Se selecciona en primer lugar la variable a partir de la cual se diferenciarán los dos grupos, en este caso *Género* y se clickea sobre la flecha (con dirección hacia la derecha). De esta manera la palabra *Género* se trasladará al recuadro *Criterio de clasificación*. Posteriormente se selecciona la variable continua de interés (en este caso A*nsiedad*), se hace click sobre la flecha para trasladar la palabra *Ansiedad* al recuadro *Variables*. Se desplegará un nuevo menú en el que podemos seleccionar algunas opciones para realizar la prueba *t* (seleccionar si utilizaremos una prueba de una o dos colas, determinar el nivel de homogeneidad requerido para realizar una corrección Satterwait, mostrar las varianzas de los grupos, entre otras opciones), finalmente debemos clickear sobre al botón *Aceptar* para ejecutar la prueba.

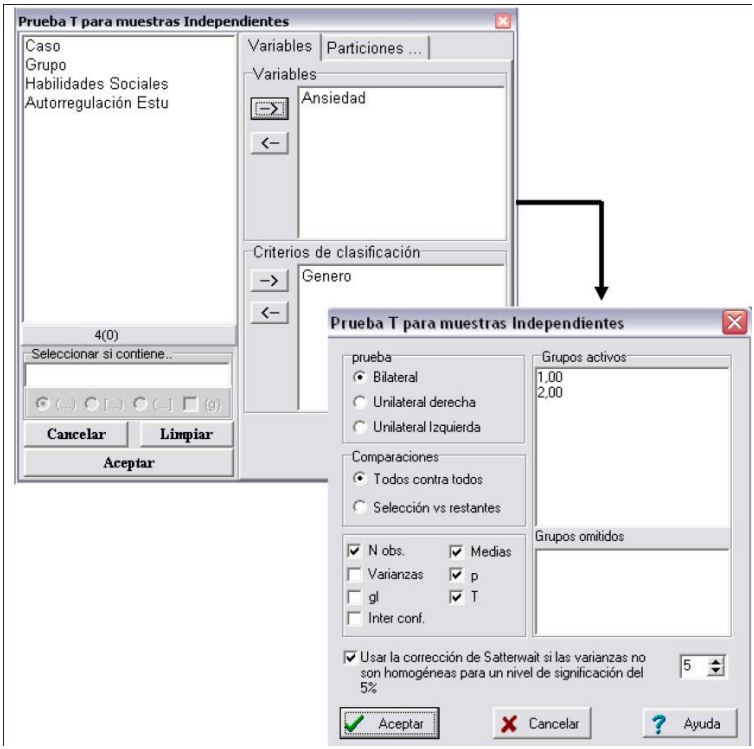


Figura 6: Calculo de Prueba t para muestras independientes con InfoStat

-Paso 3: Al clickear en *Aceptar* se desplegarán los resultados del análisis. Para realizar una correcta interpretación de los resultados se sugiere la lectura del capítulo correspondiente a "*prueba t de Student*" del presente libro.

Análisis de Varianza (ANOVA) de una Vía

Cuando pretendemos comparar más de dos grupos el procedimiento estadístico adecuado es el Análisis de Varianza (ANOVA). Si bien existen diferentes tipos de ANOVA en el presente caso nos centraremos solo sobre el ANOVA de una vía. Esta prueba permite determinar la existencia de diferencias estadísticamente significativas en una variable continua entre más de dos grupos. De esta manera esta prueba es adecuada si queremos comparar por ejemplo, si existen diferencias en los niveles de inteligencia entre personas que

poseen estudios primarios, secundarios o universitarios, o bien si la calidad de vida varía según el nivel socioeconómico (marginal, bajo, medio o alto). En el presente caso examinaremos si existen diferencias en los niveles de *Ansiedad* según si el *Grupo* recibió tratamiento (Grupo 1), no recibió tratamiento (Grupo 2) o recibió un tratamiento placebo (Grupo 3). Para ello debemos seguir los siguientes pasos:
-Paso 1: Colocar el curso en *Estadísticas*, moverlo hacia abajo y colocarlo sobre *Análisis de Varianza* (figura 7).

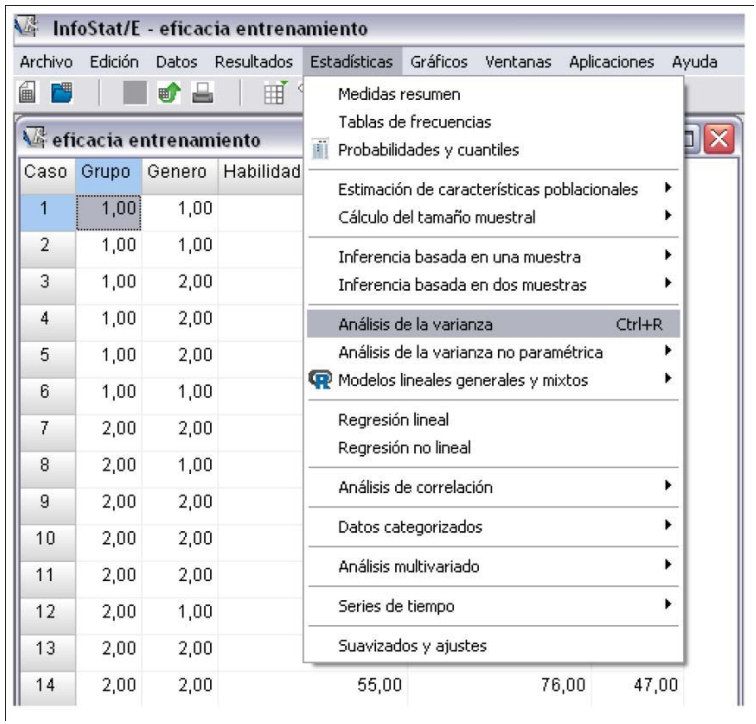


Figura 7: Calculo de ANOVA de una vía con InfoStat

-Paso 2: Se abrirá un menú que contiene las variables de la base de datos (figura 8). Se selecciona en primer lugar la variable a partir de la cual se diferenciarán los dos grupos, en este caso *Grupo* y se traslada hacia el recuadro *Criterio de clasificación* clickeando sobre la flecha correspondiente. Posteriormente se selecciona la variable continua de interés (en este caso A*nsiedad*), se hace click sobre la flecha para trasladar la palabra *Ansiedad* al recuadro *Variables*.

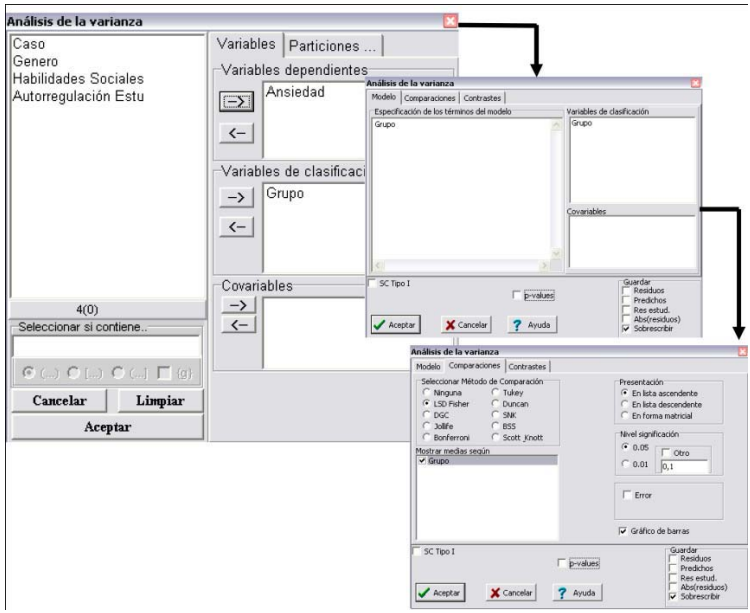


Figura 8: Calculo de ANOVA de una vía con InfoStat

-Paso 3: Se abrirá un nuevo menú en el cual pueden seleccionarse algunas opciones para realizar el análisis de varianza (figura 8). A diferencia de los menús desplegados en las anteriores pruebas estadísticas, en este caso se presentan tres solapas (*Modelo*, *Comparaciones* y *Contraste*). Especial atención merece la solapa *Comparaciones* ya que en la misma deberemos especificar el método de comparación post hoc utilizado (en el caso del ejemplo se seleccionó LSD de Fisher). Una vez seleccionado clickeamos en *Aceptar*.
-Paso 4: Se desplegará en una nueva ventana los resultados del análisis. Para realizar una correcta interpretación de los mismos se sugiere la lectura del capítulo correspondiente a "*Análisis de Varianza*" del presente libro.

Consideraciones Finales

El uso de software estadístico constituye una destreza clave para un manejo eficiente de los procedimientos estadísticos. Un conocimiento en profundidad de las fórmulas y bases matemáticas de una prueba estadística sin un manejo mínimo de los softwares requeridos para

aplicar dichas pruebas, supone un saber incompleto de la estadística. Tomando esto en consideración es que se optó por incluir el presente apéndice. Sin embargo cabe señalar algunas precauciones y sugerencias.

En primer lugar, las facilidades que brindan estos programas pueden resultar perjudiciales si el usuario no posee interiorizadas las bases lógicas y matemáticas de los procedimientos estadísticos utilizados. El software no puede juzgar si el procedimiento que seleccionamos es el adecuado para nuestro problema, así como tampoco nos informará si estamos utilizando el procedimiento acorde a los niveles de medición de las variables de nuestro estudio, por ejemplo. Es por ello que siempre que utilicemos un programa estadístico debemos hacerlo con pleno conocimiento y comprensión de los procedimientos estadísticos que pretendemos aplicar. Se sugiere la lectura de los capítulos del presente libro antes de utilizar las aplicaciones del InfoStat.

En segundo lugar debe considerarse que el InfoStat es un software muy completo y versátil, en el presente apéndice sólo se consideraron algunas de las pruebas paramétricas más utilizadas en psicología. Se sugiere la visita a la página www.infostat.com.ar para un mayor conocimiento de las aplicaciones de este programa.

En tercer y último lugar, debe considerarse que el programa InfoStat no es el único software estadístico que podemos utilizar. Por el contrario existe una gran cantidad de programas que poseen una gran variedad de aplicaciones estadísticas básicas y avanzadas. Se sugiere la lectura de la bibliografía referida para tomar conocimiento de algunos de los softwares más destacados.

Referencias Bibliográficas

Aragón, S. y Méndez, M. (2005). *Aplicaciones de la estadística a la psicología*. México: Editorial Porrúa

Aron, A. y Aron, E. N. (2001). *Estadística para Psicología*. Argentina: Pearson Education.

Balzarini M.G., Gonzalez L., Tablada M., Casanoves F., Di Rienzo J.A., Robledo C.W. (2008). *Manual del Usuario*, Editorial Brujas, Córdoba, Argentina.

Blalock, H. (1986). *Estadística social*. México: Fondo de Cultura Económica.

Clairin, R. y Brion P. (1996)*: Manuel de Sondages*. Centre Francais sur la Population et le Developpement, Paris

Cohen J. (1994). "The Earth is Round (p<.05)". *American Psychologist,* vol 49, N. 12, 997-1003

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ:Erlbaum

Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W. InfoStat versión 2010. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

Durkheim, E. (1994 [1897]). *El suicidio. Estudio de Sociología* Buenos Aires: Centro editor de América Latina.

Dziula, C., y Reyna, C. (2005). *Tolerancia rápida a los efectos del etanol en ratas periadolescentes: efectos diferenciales sobre patrones térmicos y de conducta exploratoria* (Tesis de grado no publicada). Facultad de Psicología, Universidad Nacional de Córdoba, Argentina.

Everitt S. & Wykes T. (2001). *Diccionario de estadística para psicólogos*. Barcelona: Ariel

Field A. (2005): *Discovering statistics using SPSS*: London Thousand Oaks: Sage Publications

Gardner, R. (2003*). Estadística para Psicología Usando SPSS*. México: Pearson Education.

Garret, H. (1983). *Estadística en Psicología y Educación*. Barcelona: Ediciones Paidós.

Grasso, L. (2007): *Estadística para las Ciencias Sociales y del Comportamiento*. Universidad Nacional de Córdoba: Dirección General de Publicaciones.

Guerrero R., González, C. y Medina, E. (1986). *Epidemiología*. Addison-Wesley Iberoamericana.

Hadot, P. (2006). *Ejercicios espirituales y filosofía antigua*. Madrid: Ediciones Siruela.

Herran, F. (2002). "Qu'est-ce que la démographie? Voyage historique et critique au pied des pyramides" en Université tous les savoirs *La Géographie et la Démographie*. Paris: Odile Jacob

Hite, S. (1976). *The Hite Report*. Citado por Horton, P. y Hunt, Ch.: *Sociología*. McGraw-Hill/ Interamericana de México: México, 1988

Kazdin, A. E. (2001). *Métodos de Investigación en Psicología Clínica*. México: Prentice Hall.

Kendall M. G. (1948): *Rank correlation methods*. Londres: Griffin

Ledesma, R.; Molina, G. J.; Valero-Mora, P. & Poó, F. (2007) Software Estadístico de Libre Acceso en Psicología. Una Librería de Módulos para el Sistema ViSta. *Evaluar, 7,* 19-33. Disponible en http://www.revistaevaluar.com.ar/72.pdf.

Ledesma, R.; Valero-Mora, P. & Molina, G. J. (2010). Vista: Un Software para la Enseñanza de la Estadística y la Psicometría. *Revista Argentina de Ciencias del Comportamiento, 2 (2),* 52-59. Disponible en http://www.psyche.unc.edu.ar/racc/index.php/comportamiento/article/viewFile/50/Ledesma.

Manzano, V., Varela, J., García, A. & Pérez, J.F. (1999) *SPSS para Windows*. Madrid: Ra-Ma

Manzano, V.A. & Tobio T.B. (2003) *Análisis de datos y técnicas de muestreo*. En Lévy Mangin, J. P. & Varela Mallou, J. (2003). *Análisis Multivariable para las Ciencias Sociales*. España: Prentice Hall.

Medrano, L. (2008). Utilización de Software de Libre Acceso para la Enseñanza de Estadística y Psicometría. *Revista de Enseñanza de la Psicología: Teoría y Experiencia, 4 (1),* 1-7. Disponible en:

http://psicologia.udg.es/revista/publicacions/04/Cast/01(4)_Cast.pdf

Medrano, L.; Mirantes, R. & Marchetti, P. (2009) "Evaluación del Impacto de un Programa de Entrenamiento en Aprendizaje Autorregulado y Habilidades Sociales Académicas en Ingresantes Universitarios". *Revista Argentina de Ciencias del Comportamiento 1 (2),* 116-117.

Organización Panamericana de la Salud (2003): *Informe mundial sobre la violencia y la salud*. Oficina Regional para las Américas de la Organización Mundial de la Salud, Washington, D.C.

Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos. Aplicaciones con SPSS*. Madrid: Pearson Education.

Quivy, R. y Campenhoudt, L. (2006). *Manual de Investigación en Ciencias Sociales* – Limusa

Raven, J. C. (1936). "Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive". MSc Thesis, University of London.

Ridao García I. & Gil Flores J. (2002). "La jornada escolar y el rendimiento de los alumnos" *Revista de Educación*, núm. 327 pp. 141-I56

Rodriguez, N. (1995): *Curso Taller sobre Diseño de Muestras Probabilísticas XXIII Coloquio Argentino de Estadística.* Villa Giardino, 1995

Roitter, H. (1994): *Temas Introductorios para el estudio del muestreo*. Departamento de Estadística y Matemática, Facultad de Ciencias Económicas, Universidad Nacional de Córdoba, 1994

Scheaffer, R., Mendenhall, W. y Ott, L. (1897). *Elementos de Muestreo*. Grupo Editorial Iberoamérica: México DF

Selltiz, C., Wrightsman, L., Deutsch, X. y Cook, S. (1980). *Métodos de investigación en las relaciones sociales*. Madrid: Rialp.

Siegel S. (1956): *Nonparametric Statistics for the Behavioral Sciences* Nueva York: McGraw-Hill,

Spearman, C. (1904). "General intelligence objectively determined and measured". *American Journal of Psychology*, 15, 201–293.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition*. London: Macmillan.

Stevens, S. (1951). "Mathematics, measurement and psychophysics". En S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Tankard, J. (1984). *The Statistics Pioneers* Cambridge Massachusetts: Schenkman

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Índice

Impreso por Editorial Brujas
abril de 2011
Córdoba - Argentina