

Multiple choice questions

1. Which of the following are true statements?

- I. High correlation does not necessarily imply causation.
- II. A lurking variable is a name given to variables that cannot be identified or explained.
- III. A pattern in the residual plot indicates that a lurking variable is involved.

- (A) I only
- (B) II only
- (C) III only
- (D) I and II only
- (E) I and III only

2. All but one of these statement contains a blunder. Which could be true?

- (A) The correlation between a football player's weight and the position he plays is 0.54.
- (B) The correlation between the amount of fertilizer used and the yield of beans is 0.42.
- (C) The correlation between a car's length and its fuel efficiency is 0.71 miles per gallon.
- (D) There is a high correlation (1.09) between height of a corn stalk and its age in weeks.
- (E) There is a correlation of 0.63 between gender and political party.

3. Two variables that are actually not related to each other may nonetheless have a very high correlation because they both result from some other, possibly hidden factor. This is an example of

- (A) leverage.
- (B) a confounding variable.
- (C) extrapolation.
- (D) a common response variable.
- (E) an outlier.

4. Which is true?

- I. Random scatter in the residuals indicates a model with high predictive power.
- II. If two variables are very strongly associated, then the correlation between them will be near +1.0 or -1.0.
- III. The higher the correlation between two variables the more likely the association is based on cause and effect.

- (A) I only
- (B) II only
- (C) I and II only
- (D) I, II, and III
- (E) None

Survey Vocabulary – you should understand each of the following terms:

Population	Sample	Bias	Census
Population parameter	Sample statistic	Representative sample	Simple random sample (SRS)
Stratified random sample	Cluster sample	Multistage sample	Systematic sample
Voluntary response bias	Convenience sample	Nonresponse bias	Undercoverage

Free response questions

5. **Birth rates.** The table shows the number of live births per 1000 women aged 16–44 years in the United States in 1965 (*National Vital Statistics Report, April 2001*).

Year	1965	1970	1975	1980	1985	1990	1995	2000
Rate	19.4	18.4	14.8	15.9	15.6	16.4	14.8	14.5

- Make a scatterplot and describe the general trend in *birth rates*. (Enter Year as 1965, 1970, etc.)
 - Find the equation of the regression line.
 - Check to see if the line is an appropriate model. Explain.
 - Interpret the slope of this line.
 - The table gives rates only at 5-year intervals. Estimate what the rate was in 1978.
 - In 1978 the birth rate was actually 15.0. How close did your model come?
 - Predict what the *birth rate* will be in 2005. Comment on your faith in this prediction.
 - Predict the *birth rate* for 2020. Comment on your faith in this prediction.
6. **Baseball salaries.** Ball players have been signing ever larger contracts. The highest salaries (in millions of dollars per season) for some notable players are given in the following table.

Player	Year	Salary (million \$)	Player	Year	Salary (million \$)
Nolan Ryan	1980	1	Pedro Martinez	1998	12.5
George Foster	1982	2.04	Mike Piazza	1999	12.5
Kirby Puckett	1990	3	Mo Vaughn	1999	13.3
Jose Canseco	1990	4.7	Kevin Brown	1999	15
Roger Clemens	1991	5.3	Carlos Delgado	2001	17
Ken Griffey Jr.	1996	8.5	Alex Rodriguez	2001	25.2
Albert Belle	1997	11			

- Plot the data and comment on the scatterplot.
 - Re-express the data using an exponential transformation to straighten the scatterplot.
 - Create an exponential model for the trend in salaries.
 - What does the R^2 of your model mean in the context of the problem.
 - Predict a superstar salary for 2005.
 - Create a power-law model for the trend in salaries
 - Use the power-law model to predict a superstar salary in 2005.
7. **Louisiana Forests.** To gather data on a 1200-acre pine forest in Louisiana, the US Forest Service laid a grid of 1410 equally spaced circular plots over a map of the forest. A ground survey will visit a simple random sample of 10% of these plots.
- How would you number the plots in order to take the simple random sample?
 - Use the table of random digits beginning below to choose the first 2 plots in an SRS of 141 plots.

95592	94007	69971	91481	60779	53791	17297	59335
68417	35013	15529	72765	85089	57067	50211	47487

8. **El Niño.** Concern over the weather associated with El Niño has increased interest in the possibility that the climate on earth is getting warmer. The most common theory relates an increase in atmospheric levels of carbon dioxide (CO₂), a greenhouse gas, to increases in temperature. A regression was performed on the mean annual CO₂ concentration in the atmosphere measured in parts per million (ppm) at the top of Mauna Loa in Hawaii and the mean annual air temperature over both land and sea across the globe in degrees Celsius (C). The regression output for predicting *temperature* from CO₂ levels produces the following output table.

Dependent variable is: Temperature

R-squared = 33.4%

Variable	Coefficient
intercept	15.3066
CO ₂	0.004

- What is the correlation between CO₂ level and temperature?
 - Explain the meaning of R^2 in this context
 - Give the regression equation.
 - What is the meaning of the slope in this equation?
 - What is the meaning of the intercept in this equation?
9. **Cars.** A survey of autos parked in the student and staff lots at a large university classified by the cars brand's country of origin are shown in the table below.

	Student	Staff	Total
American	107	105	
European	33	12	
Asian	55	47	
Total			

- Calculate all of the row and column totals for this table.
 - What percent of all cars are foreign (non-American)?
 - What percent of American cars were owned by students?
 - What percent of students owned American cars?
 - Show that these two categorical variables are not independent.
10. **New drugs.** You have designed a new pain relieving drug (Wonder Drug) that initial evidence suggests works much faster than ibuprofen (Advil and Motrin) at reducing fever in children. You have 100 children (volunteered by their parents) in a clinic all with fevers. Design an experiment that compares ibuprofen at the standard dose to the new drug at three doses (1 mg/kg, 2 mg/kg, and 3mg/kg). Be sure to follow the principles of good experimental design. Also, identify the factors, levels, treatments, and explanatory and response variables.
11. **Unusual points in scatter plot.** What are the three types of unusual points that can be found in scatter plots between two quantitative variables? How do these three types of points affect the strength of the correlation and the parameters of a linear model? How big of a residual do each of the types of points have? What is leverage in this context? What gives a point leverage?
12. **Lurking variables.** What are the two types of lurking variables that we have discussed in class? Give an example of each. What type of study is required to prove a cause and effect relationship between two or more variables? What type of study although useful cannot prove cause and effect?