Name Key 2018 Period

AP Statistics | Bivariate Data Analysis Test Review

Multiple-Choice

1. The correlation coefficient measures:
   (a) Whether there is a relationship between two variables
   (b) The strength of the relationship between two quantitative variables
   (c) Whether or not a scatterplot shows an interesting pattern
   (d) Whether a cause and effect relation exists between two variables
   (e) The strength of the linear relationship between two quantitative variables

E

2. Which of the following is true of the correlation r?
   (a) It is a resistant measure of association
   (b) $-1 \leq r \leq 1$
   (c) If r is the correlation between X and Y, then –r is the correlations between Y and X
   (d) Whenever all the data lie on a perfectly straight-line, the correlations r will always be equal to +1.0
   (e) All of the above

B

3. A study gathers data on the outside temperature during the winter in degrees Fahrenheit and the amount of natural gas a household consumes in cubic feet per day. Call the temperature x and gas consumption y. The house is heated with gas, so x helps explain y. The least squares regression line for predicting y from x is: $\hat{y} = 1344 - 19x$. When the temperature goes up 1 degree, what happens to the gas usage predicted by the regression line?
   (a) It goes up 19 cubic feet
   (b) It goes down 19 cubic feet
   (c) It goes up 1344 cubic feet
   (d) It goes down 1344 cubic feet
   (e) Can't tell without seeing the data

B

4. A least squares regression line is created for predicting stopping distance (y) in feet from temperature (x) in degree Fahrenheit. If stopping distance was expressed in yards instead of feet, how would the correlation r between temperatures and stopping distance change?
   (a) r would be divided by 12
   (b) r would be divided by 3
   (c) r would not change
   (d) r would be multiplied by 3
   (e) r would be multiplied by 12

C

5. The fraction of the variation in the values of y that is explained by the least-squares regression of y on x is...
   (a) The correlation coefficient
   (b) The slope of the least-squares regression line
   (c) The square of the correlation coefficient
   (d) The intercept of the least-squares regression line
   (e) The slope of a transformed exponential function

C

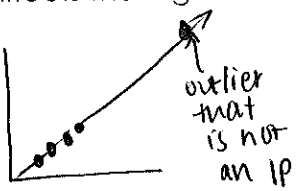
6. Volunteers for a research study were divided into three groups. Group 1 listened to Western religious music, group 2 listened to Western rock music, and group 3 listened to Chinese religious music. The blood pressure of each volunteer was measured before and after listening to the music, and the change in blood pressure (blood pressure before listening minus blood pressure after listening) was recorded. To explore the relationship between type of music and change in blood pressure, we could:

C

(a) See if blood pressure decreases as type of music increases by examining a scatterplot
(b) Make a histogram of the change in blood pressure for all the volunteers
(c) Make a side-by-side boxplots of the change in blood pressure, with a separate boxplot for each group
(d) Do all of the above
(e) Do none of the above

7. Which of the following statements about influential points are true?

I. Influential points have large residuals not if they pull the LSRL close enough
II. Removal of an influential point sharply affects the regression line
III. Outliers are always influential points *



can, but it doesn't have to necessarily be sharply

E

(a) I and II
(b) I and III
(c) II and III
(d) I, II, and III
(e) None of the above gives the complete set of true responses.

8. Which of the following statements concerning residuals is true?

D

(a) The sum of the residuals is always 0
(b) A plot of the residuals is useful for assessing the fit of the least squares regression line
(c) The value of the residual is the observed value of the response minus the value of the response that one would predict from the least-squares regression line
(d) All of the above
(e) None of the above

9. Using least squares regression, I determine that the logarithm (base 10) of the population of a country is approximately described by the equation: $\log(\widehat{population}) = -13.5 + 0.01(year)$. Based on this equation, the population of the country in the year 2000 should be about:

D

(a) 6.5
(b) 665
(c) 2,000,000
(d) 3,167,277
(e) None of the above

10. Which of the following would provide evidence that a power law model appropriately describes the relationship between a response y and an explanatory variable x?

D

(a) A scatterplot of y versus x looks approximately linear
(b) A scatterplot of log y versus x looks approximately linear
(c) A scatterplot of y versus log x looks approximately linear
(d) A scatterplot of log y versus log x looks approximately linear
(e) None of the above

Free-Response

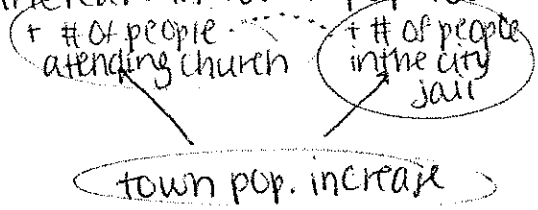
11. Over the past few years in a small town, there has been a strong positive correlation between the number of people attending church and the number of people in the city jail.

(a) Is this evidence that going to church causes people to go to jail? Explain.

No. correlation ≠ causation. There could be other factors involved (hidden variables) that cause both.

(b) What hidden variable(s) might be influencing the situation? What is this effect called? Draw a picture.

This is called common response.
an increase in town population may cause both.



12. Data is collected from 189 elementary school students about their favorite subject in school and their gender. The data is displayed in the table below.

	Gender			
Subject	Male	Female	**Total**	%
Math	23	26	49	25.9%
Science	24	23	47	24.9%
Reading	15	23	38	20.1%
PE	28	27	55	29.1%
Total	90	99	189	100%
%	47.6%	52.4%	100%	

(a) Compute the marginal distribution of favorite subject for all students.

subject	All students
math	49/189 = 25.9%
science	47/189 = 24.9%
reading	38/189 = 20.1%
PE	55/189 = 29.1%
total	189/189 = 100%

(b) Compute the conditional distributions of gender among students whose favorite subject is reading.

gender	reading
male	15/38 = 39.5%
female	23/38 = 60.5%
total	38/38 = 100%

(c) What percentage of male students have math as their favorite subject? 23/90 = 25.6%.
25.6% of male students have math as their favorite subject.

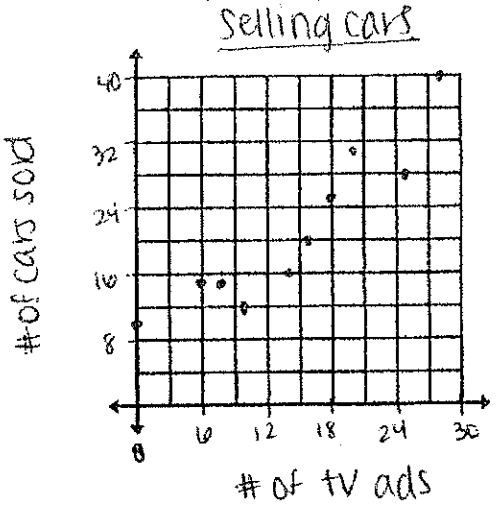
(d) What percentage of female students have science as their favorite subject? 23/99 = 23.2%
23.2% of female students have science as their favorite subject.

(e) What percentage of students whose favorite subject is PE are female?
27/55 = 49.1% of students whose favorite subject is PE are female.

13. A local car dealership has been using 1-minute spot ads on a local TV station. The ads always occur during the evening hours and advertise the different models and price ranges of cars on the lot that week. During a 10-week period, the dealer kept a weekly record of the number of TV ads and the number of cars sold. The results are shown in the table below.

TV ads	6	20	0	14	25	16	28	18	10	8
Cars sold	15	31	10	16	28	20	40	25	12	15

(a) Construct a scatterplot of the data and describe the relationship between TV ads and cars sold. Which is the explanatory variable and which is the response variable?



explanatory variable: # of tv ads per week

response variable: # of cars sold per week

There is a weak, positive, linear relationship between # of tv ads per week and # of cars sold per week. The scatterplot shows no large gaps and no clusters in the data.

(b) Construct a linear regression model for the data. Is the model a good fit for the data? Use statistical data to justify your response.

$\widehat{cars} = 6.541 + 1.011(tv\ ads)$

This is a good fit because the relationship between # of ads and # of cars sold is fairly linear and the correlation coefficient $r = 0.919$ which is fairly close to 1.

(c) The manager decided that they can only afford 12 ads per week. At that level of advertisement, how many cars can the dealership expect to sell?

$\widehat{cars} = 6.541 + 1.011(12) = 18.67$ cars

The dealership can expect to sell about 18 cars per week if they run 12 ads per week.

(d) The next week there were 22 ads that were shown. According to this model, its residual for number of ads is 3. How many cars were sold? Round to the nearest whole car if necessary.

$residual = y - \hat{y}$

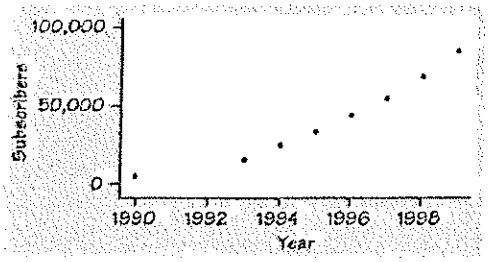
$3 = y - 28.78$

$y = 31.78$

$\hat{y} = 6.541 + 1.011(22) = 28.78$

The dealership sold about 31 cars in a week when 22 ads were shown.

14. Cell phones have revolutionized the way we do business and the way we stay in touch with friends and family. The cell phone industry enjoyed substantial growth in the 1990s. The scatterplot below shows the number of cell phone subscribers (in the thousands) in the United States between 1990 and 1999.



(a) Interpret the relationship between year and the number of subscribers.

There is a strong, positive, slightly curved relationship between year and # of cell phone subscribers. There is a small gap in data between the year 1990 and 1993. (on the scatterplot)

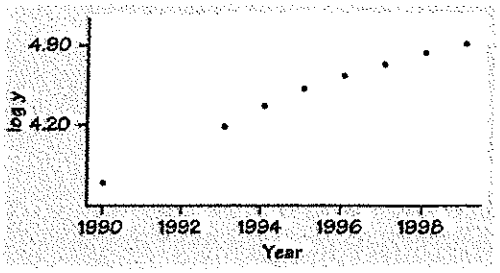
Researchers transformed the number of subscribers using logarithms. Applying least squares regression to the transformed data gave the following computer output, plot of transformed values, and residual plot:

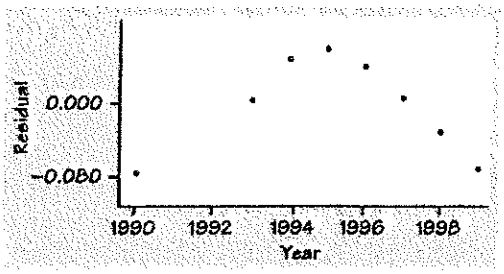
a b

$\widehat{log(subscribers)} = -263 + 0.134(year)$

Predictor		Coef	Stdev	t-ratio	p
Constant	a	-263.20	14.63	-17.99	0.000
year	b	0.134170	0.007331	18.30	0.000

s = 0.05655 R-sq = 98.2% R-sq(adj) = 97.9%





(b) Perform the inverse transformation to obtain a non-linear model for the original data.

$\widehat{Subscribers} = 10^{-263.20} \cdot 1.3614^{(year)}$

(c) Based on the information above, is this model appropriate for making predictions about the number of cell phone subscribers based on year? Explain.

This model is not appropriate. Although $r = 0.9910$ is close to 1, the residual plot shows a clear pattern, which means the model is not ideal. (curve)

(d) Interpret R^2 in context.

98.2% of the variation in the # of cellphone subscribers can be accounted for by the LSRL of # of cell phone subscribers on year.

15. Each year, students in an elementary school take a standardized math test at the end of the school year. For a class of 4th graders, the average score was 55.1 with a standard deviation of 12.3. In the 3rd grade, these same students had an average score of 61.7 with a standard deviation of 14.0. The correlation between the two sets of scores is r = 0.95.

(a) Calculate the equation of the least-squares regression line for predicting a 4th (y) grade score from a 3rd (x) grade score. SHOW YOUR WORK!

$\bar{x} = 61.7$ $\quad S_x = 14.0$

$\bar{y} = 55.1$ $\quad S_y = 12.3$

$r = 0.95$

$\hat{y} = a + bx$

$b = r\frac{S_y}{S_x} = (0.95)\frac{12.3}{14.0} = 0.8346$

$a = \bar{y} - b\bar{x} = 55.1 - 0.8346(61.7)$

$= 3.1603$

(4th grade score) = 3.1603 + 0.834(3rd grade score)

(b) Calculate R^2 and explain it in terms of the context of the problem.

$r = 0.95$ $\quad R^2 = 0.9025$

90.25% of the variation in the values of 4th grade math scores is accounted for by the LSRL of 4th grade math scores on 3rd grade math scores.

(c) Can this equation be used to represent the student's scores when they are in 12th grade? Why or why not and explain your answer?

NO, because this data only represents the variables 3rd and 4th grade math scores. 12th grade math scores would be a new variable and we would need a new data set and LSRL to predict those.