AP Stats
Review for the AP Exam - Chapters 3 & 4

Name: Key
Hour: _____

Chapter 3:

1. What is the explanatory variable? The response variable?

Explanatory - x Response - y

2. How do you describe a scatterplot?

T Trend (+/-)
U unusual features
S shape (linear, curved, etc)
S strength weak, moderate, strong

3. What are outliers and influential points in a scatterplot?

Outliers lie beyond $Q_3 + 1.5IQR$ or $Q_1 - 1.5IQR$
Influential points skew the LSR closer to the influential point

4. What is the correlation coefficient and what does it measure?

r - Measures the strength and direction of the linear relationship between x and y.

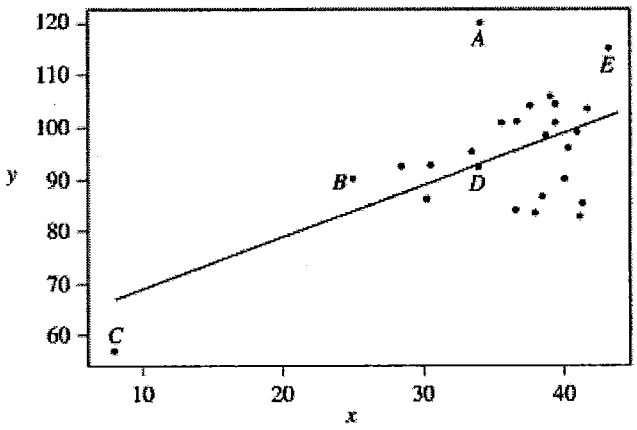
5. What is the coefficient of determination and what does it measure?

r^2 - measures the percent of the variation in y that can be explained by the linear relationship between x and y.

6. What is extrapolation? Should you do it?

Do not use extrapolation. Extrapolation makes predictions for x values outside the range of the collected data.

7. Remember: Correlation does NOT equal causation.



8. In the scatterplot of *x* versus *y* shown above, the least squares regression line is superimposed on the plot. Which of the following points has the largest residual?

(A) A (B) B (C) C (D) D (E) E

9. The computer output below shows the result of a linear regression analysis for predicting the concentration of zinc, in parts per million (ppm), from the concentration of lead, in ppm, found in fish from a certain river.

Response variable is Zinc (ppm)

Variable	Coefficient	Std Dev	T	P
Constant	16.3	4.90	3.32	0.003
Lead (ppm)	19.0	1.69	10.01	0.000

S = 16.17 R-Sq = 82.0%

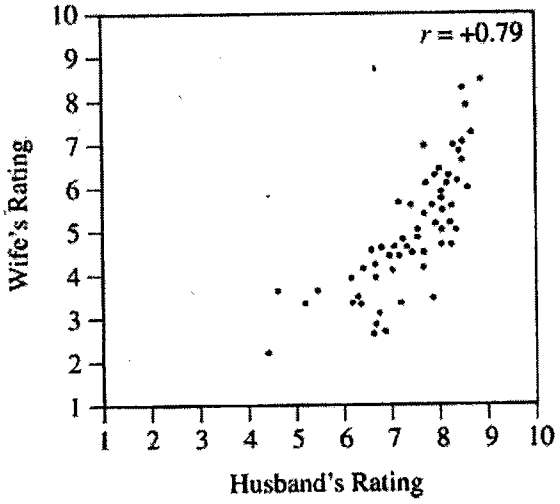
Which of the following statements is a correct interpretation of the value 19.0 in the output?

(A) On average there is a predicted increase of 19.0 ppm in concentration of lead for every increase of 1 ppm in concentration of zinc found in the fish.
(B) On average there is a predicted increase of 19.0 ppm in concentration of zinc for every increase of 1 ppm in concentration of lead found in the fish.
(C) The predicted concentration of zinc is 19.0 ppm in fish with no concentration of lead.
(D) The predicted concentration of lead is 19.0 ppm in fish with no concentration of zinc.
(E) Approximately 19% of the variability in zinc concentration is predicted by its linear relationship with lead concentration.

10. The correlation between two values X and Y equals 0.8. If both the X values and the Y values are converted to z-scores, then the correlation between the z -scores for X and the z-scores for Y would be

(A) –0.8 (B) –0.2 (C) 0.0 (D) 0.2 (E) 0.8

11. In a recent survey, randomly selected married couples from the same town were asked to rate the overall quality of living in their town on a scale from (very poor) to (excellent) on twenty different attributes such as accessibility to major highways, availability of entertainment, services provided by tax dollars, etc. For each couple, the husband's individual ratings on the twenty attributes were averaged to produce an overall quality rating, and that process was repeated for the wife. Each point on the scatterplot below displays the overall rating of one of the couples with the husband's rating represented by the horizontal axis and the wife's rating represented by the vertical axis.



Based on the scatterplot, which of the following statements is true?

(A) Husbands tended to rate the quality of living higher than their wives did.
(B) More overall ratings of 7 or less were assigned by husbands than by wives.
(C) The range in the husbands' overall ratings is greater than the range in the wives' overall ratings.
(D) The difference in overall ratings between a husband and wife was not more than 3 for any couple.
(E) For each couple, the overall rating assigned by the husband was the same as the overall rating assigned by the wife.

12. A least squares regression line was fitted to the weights (in pounds) *versus* age (in months) of a group of many young children. The equation of the line is $\hat{y} = 16.6 + 0.65t$,

$\hat{y} = 16.6 + .65(20) = 29.6$ $\quad y - \hat{y} = 25 - 29.6 = -4.6$

where $\hat{y}$ is the predicted weight and t is the age of the child. A 20-month-old child in this group has an actual weight of 25 pounds. Which of the following is the residual weight, in pounds, for this child?

(A) –7.85 (B) –4.60 (C) 4.60 (D) 5.00 (E) 7.85

13. Each of 100 laboratory rats has available both plain water and a mixture of water and caffeine in their cages. After 24 hours, two measures are recorded for each rat: the amount of caffeine the rat consumed, *X*, and the rat's blood pressure, *Y*. The correlation between *X* and *Y* was 0.428. Which of the following conclusions is justified on the basis of this study?

(A) The correlation between *X* and *Y* in the population of rats is also 0.428.

(B) If the rats stop drinking the water/caffeine mixture, this would cause a reduction in their blood pressure.

(C) About 18 percent of the variation in blood pressure can be explained by a linear relationship between blood pressure and caffeine consumed. $(.428)^2 = .183$

(D) Rats with lower blood pressure do not like the water/caffeine mixture as much as do rats with higher blood pressure.

(E) Since the correlation is not very high, the relationship between the amount of caffeine consumed and blood pressure is not linear.

Use the information below to answer #14-17.

As part of a class project at a large university, Amber selected a random sample of 12 students in her major field of study. All students in the sample were asked to report their number of hours spent studying for the final exam and their score on the final exam. A regression analysis on the data produced the following partial computer output.

Predictor	Coef	SE Coef	T	P
Constant	62.328	4.570	13.64	0.000
Study Hours	2.697	0.745	3.62	0.005
S = 5.505	R-sq = 56.7%			

14. What is the equation of the least squares regression line?

X = study hours

$\hat{y}$ = predicted final exam scores $\quad \hat{y} = 62.328 + 2.697x$

15. Interpret the slope in context.

For each additional hour studying, the expected final exam score increases by 2.697 on average.

16. Interpret the *y*-intercept in context.

With no hours spent studying, the expected final exam score is 62.328.

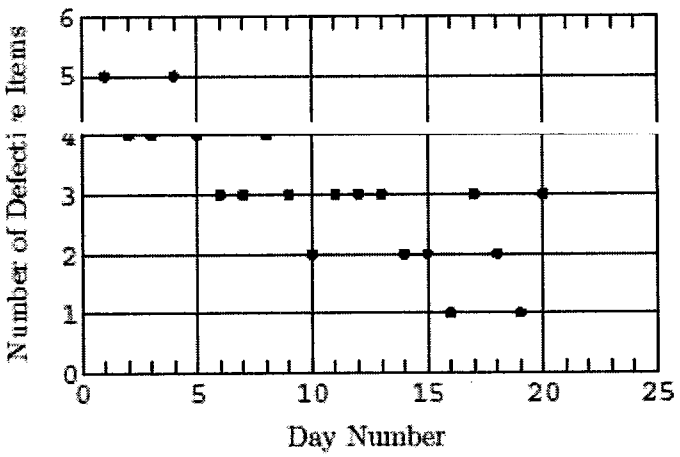
17. Find and interpret *r* in context.

$r = \sqrt{.567} = .753$

There is a moderate positive linear relationship between the # of hours spent studying and the final exam score.

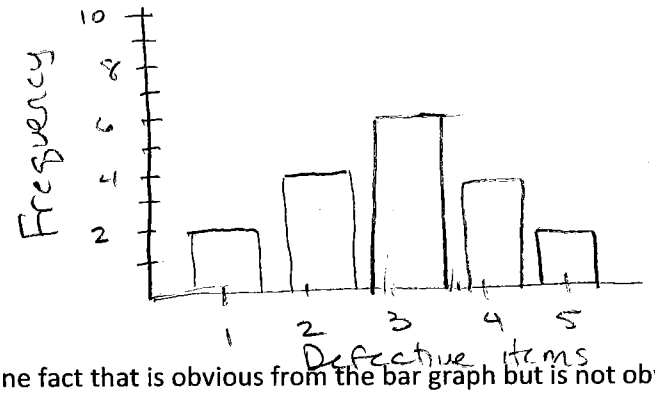
(1998 #2)

18. A plot of the number of defective items produced during 20 consecutive days at a factory is shown below.



(a) Draw a bar graph that shows the frequencies of the number of defective items.

Defective Items Produced



(b) Give one fact that is obvious from the bar graph but is not obvious from the scatterplot.

The distribution is approximately normal

(c) Give one fact that is obvious from the scatterplot but is not obvious from the bar graph.

The # of defective items has been decreasing over time

Chapter 4:

19. What transformations are used to achieve linearity in a scatterplot?

Power – ln or log of x's and y's

Exponential – ln or log of y's

Be sure to include the transformations in all statements in context!!!

20. What does it mean if two variables are confounded?

Effects of the variables on the response variable cannot be separated.

21. A local company is interested in supporting environmentally friendly initiatives such as carpooling among employees. The company surveyed all of the 200 employees at the downtown offices. Employees responded as to whether or not they own a car and to the location of the home where they live. The results are shown in the table below.

		Location of Home			
		Downtown Area In the City	Elsewhere In the City	Outside the City	Total
Car Ownership	Yes	10	15	35	60
	No	60	55	25	140
	Total	70	70	60	200

Which of the following statements about a randomly chosen person from these 200 employees is true?

(A) If the person owns a car, he or she is more likely to live elsewhere in the city than to live in the downtown area in the city. elsewhere = 15/60 downtown = 10/60

(B) If the person does not own a car, he or she is more likely to live outside the city than to live in the city (downtown area or elsewhere). outside = 25/140 downtown 115/140

(C) The person is more likely to own a car if he or she lives in the city (downtown area or elsewhere) than if he or she lives outside the city. in city 25/140 outside city 35/60 downtown 70/200

(D) The person is more likely to live in the downtown area in the city than elsewhere in the city. elsewhere = 70/200

(E) The person is more likely to own a car than not to own a car.

car = $\frac{60}{200}$ no car = $\frac{140}{200}$

22. The director of a technical school was curious about whether there is a relationship between students who complete one of the school's most popular health sciences certificate programs and whether those students go on to complete more advanced studies in the health sciences within two years of completing the certificate program. She randomly selected 100 students who completed the program. Data collected on these students are shown in the table below.

		Completed More Advanced Studies		
		Yes	No	Total
Completed Most Popular Health Sciences Certificate Program	Yes	35	25	60
	No	5	35	40
	Total	40	60	100

Which of the following statements is true for these 100 students?

(A) Being a person who completed more advanced studies is more likely than being a person who did not complete more advanced studies. adv = 40/100 no adv = 60/100 Program = 60/100 np = (40/100)

(B) Being a person who completed the program is less likely than being a person who did not complete the program.

(C) Being a person who completed the program and completed more advanced studies is less likely than being a person who did not complete the program and did not complete more advanced studies. prg ∩ adv = 35/100 np ∩ na = 35/100

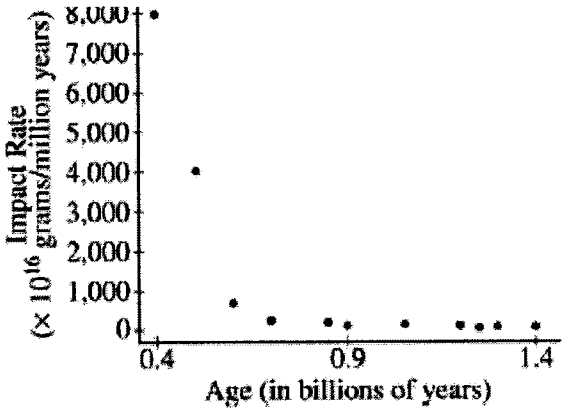
(D) Being a person who did not complete the program but completed more advanced studies is less likely than being a person who completed the program and completed more advanced studies. np ∩ adv = 5/100 p ∩ adv = 35/100

(E) Being a person who completed the program but did not complete more advanced studies is more likely than being a person who did not complete the program and did not complete more advanced studies.

program ∩ no adv = 25/100 np ∩ no adv = 35/100

(2004B #1)

23. The Earth's Moon has many impact craters that were created when the inner solar system was subjected to heavy bombardment of small celestial bodies. Scientists studied impact craters on the Moon to determine whether there was any relationship between the age of the craters (based on radioactive dating of lunar rocks) and the impact rate (as deduced from the density of the craters). The data are displayed in the scatterplot below.



(a) Describe the nature of the relationship between impact rate and age.

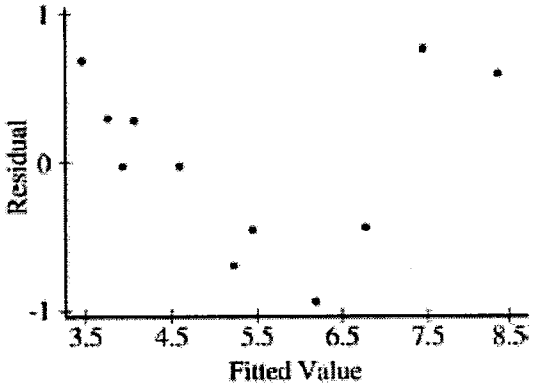
This could be exponential decay. There is a strong, negative, non-linear relationship. The impact rate drops drastically in the beginning then levels out gradually decreasing.

Prior to fitting a linear regression model, the researchers transformed both impact rate and age by using logarithms. The following computer output and residual plot were produced.

Regression Equation: ln(rate) = 4.82 – 3.92 ln(age)

Predictor	Coef	SE Coef	T	P
Constant	4.8247	0.1931	24.98	0.000
ln(age)	–3.9232	0.4514	–8.69	0.000

S = 0.5977 R-Sq = 89.4% R-Sq (adj) = 88.2%



(b) Interpret the value of r^2.

89.4% of the variation in the ln of the impact rate can be explained by the linear relationship between ln age and ln of impact rate.

(c) Comment on the appropriateness of this linear regression for modeling the relationship between the transformed variables.

The curve in the residual plot shows the linear regression model is still not appropriate.