AP Stats
Review HW #4

MULTIPLE CHOICE

1. A study was conducted of the relationship between the number of *hours* of television a student watched in the 24-hour period before a statistics examination and the *score* on the exam. The following is a computer printout from a least-square regression analysis.

Predictor	Coef	Stdev	t-ratio	p
Constant	93.052	3.426	27.16	0.000
Hours	-3.2319	0.7819	-4.13	0.001
s = 7.843	R-sq = 55.0%		R-sq (adj) = 51.7%	

Which of the following gives the correct value and interpretation of the correlation coefficient for the linear relationship between the test score and the number of hours of television watched?

A. Correlation = -0.742. The linear relationship between the test score and the number of hours of TV watched is moderate and negative.
B. Correlation = 0.550. Fifty-five percent of the variation in test scores is explained by the number of hours of TV watched.
C. Correlation = 7.416. There is a relatively weak linear relationship between the test score and the number of hours of TV watched.
D. Correlation = 0.742. About 74% of the data points lie on the least square regression line.
E. Correlation = -0.742. For every additional hour of television watched, the average test score dropped by about three-fourths of a point.

2. Allison and her twin sister Brenda both take Advanced Placement European History, Allison is in the morning class while Brenda is in the afternoon class. On the final exam, Allison received a score of 89 where the scores had a mean of 87 and a standard deviation of 3. Brenda received a score of 90 on the same test except that her class core had a mean of 89 with a standard deviation of 5. Which statement is true concerning the scores of the sisters relative to the scores on their own classes?

A. Brenda had a higher score on the test and therefore performed better relative to her own classmate's score.
B. Allison performed better relative to her own classmate's scores.
C. Allison and Brenda performed equally as well relative to their own classmate's scores.
D. We cannot compare their scores since they are in different classes.
E. We do not know how many students are in each class, so any comparison may not be fair.

3. The maker of printer cartridges for laser printers wants to estimate the mean number of documents μ that can be printed on a new high-speed printer. The company decides to test the cartridge on two dozen different laser printers. Each document is identical in number of words and amount of graphics. A histogram of the number of pages printed for each printer shows no outliers and is fairly bell-shaped. The mean and standard deviation of the sample were 3,875 sheets, and 170 sheets, respectively. It can be assumed that the laser printers were a random sample of all laser printers on the market. Which of the following is the correct formula for a 90% confidence interval for the mean number of pages printed with the new type of cartridge?

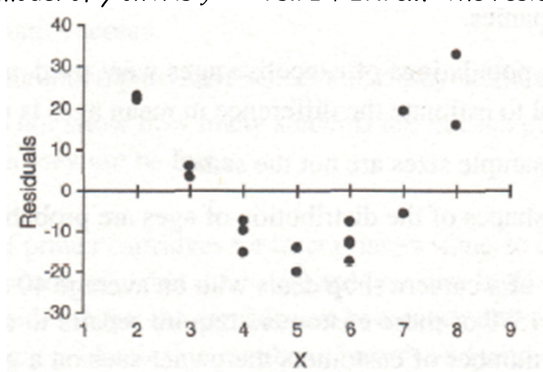
A. $3{,}875 \pm 1.711 \times \frac{170}{\sqrt{24}}$
B. $3{,}875 \pm 1.714 \times \frac{170}{\sqrt{24}}$
C. $3{,}875 \pm 1.711 \times \frac{170}{\sqrt{23}}$
D. $3{,}875 \pm 1.714 \times \frac{170}{\sqrt{23}}$
E. The company should only compute a 95% confidence interval for these data.

4. The manager at an employment agency is interested in knowing if there is a significant difference in the mean ages of executives at two rival computer software companies. This information will help him to best place people in positions at these companies. He was allowed access to the ages of all of the executives. He found a 95% *t*-confidence interval for the difference in the mean ages of all the executives at both companies. There are 86 executives at one company and 79 at the other. Why is the information provided by this interval NOT useful for this situation?
   A. There is too much variation in the length of time the workers have been at their jobs.
   B. There is too much variation in the ages of the executives at both companies.
   C. Both populations of executive ages were used so a confidence interval to estimate the difference in mean ages is not necessary.
   D. The sample sizes are not the same.
   E. The shapes of the distribution of ages are probably not normal.

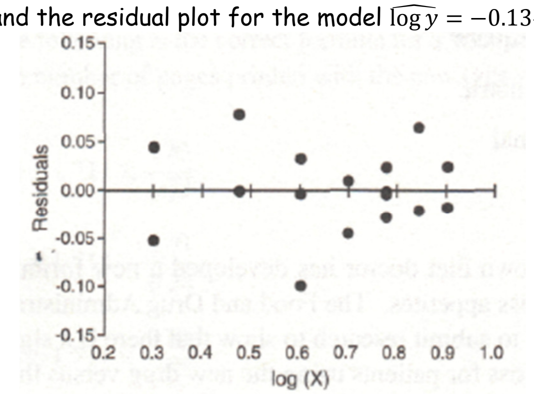
5. The owner of a camera shop deals with on average 40 customers per day. Typically, 15% of these customers require repairs to camera equipment. If X is the number of customers the owner sees on a given day until one needs repairs, which of the following best describes the probability distribution of X?
   A. Binomial
   B. Chi-square
   C. Geometric
   D. Normal
   E. *t*

6. A well-known diet doctor has developed a new formula for a drug that will suppress appetites. The Food and Drug Administration (FDA) wants this doctor to submit research to show that there is a significant difference in weight loss for patients using the new drug versus the old drug. Which of the following would be the best method to obtain results the FDA is requesting?
   A. The doctor should randomly choose patients and allow them to select which drug they want to use.
   B. The doctor should randomly choose patients to take the new drug and ask patients using the old drug how much weight they have lost.
   C. The doctor should randomly assign patients to two groups: one group takes the new drug and the other group takes the old drug.
   D. The doctor should randomly assign the patients to two groups: one group takes the new drug and the other group takes a placebo.
   E. The doctor assigns the new drug to patients who have more than 10 pounds to lose and the old drug to patients with less than 10 pounds to lose.

7. A least-square regression model of y on x is $\hat{y} = -73.71 + 27.76x$. The residual graph is shown below.



A transformation is made and the residual plot for the model $\widehat{\log y} = -0.1348 + 2.623 \log x$ is shown.



Which of the following is the correct conclusion given the graphs above?

A. Neither model provides an appropriate relationship between the variables.
B. There is a linear relationship between the two variables; the first model is more appropriate.
C. There is a linear relationship between the two variables; the second model is more appropriate.
D. There is a nonlinear relationship between the two variables; the first model is more appropriate.
E. There is a nonlinear relationship between the two variables; the second model is more appropriate.

8. In a test of H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 > \mu_2$, samples from approximately normal populations produces means of $\bar{x}_1 = 24.3$ and $\bar{x}_2 = 22.4$. The *t*-test statistic is 1.44 and the *P*-value is 0.085. Based upon these results, which of the following conclusions can be drawn?
   A. There is a good reason to conclude that $\mu_1 > \mu_2$ because $\bar{x}_1 > \bar{x}_2$.
   B. About 8.5% of the time $\bar{x}_1 = \bar{x}_2$.
   C. About 8.5% of the time $\mu_1 = \mu_2$.
   D. The null hypothesis can be rejected since $\bar{x}_1 - \bar{x}_2$ is greater than the *t*-test statistic.
   E. The observed difference between $\bar{x}_1$ and $\bar{x}_2$ is significant at the α = 0.10 level.

9. X and Y are independent random variables. X is normally distributed with mean 100 and standard deviation 8. Y is normally distributed with mean 96 and standard deviation 6. For randomly generated values of X and Y, what is the probability that X is greater than Y?
   A. 0.0401
   B. 0.6554
   C. 0.8273
   D. 0.9772
   E. 1

10. A student studying the commuting habits of students at a particular college wants to compare the mean commuting time for undergraduates with that of graduate students. The question is whether or not there is a significant difference between the mean travel time. Random samples of size 50 are taken from each group and their commuting times recorded. Which would be the appropriate test procedure to use in this situation?
   A. One-sample *z*-test
   B. Two-sample *z*-test
   C. Paired *t*-test
   D. Two-sample *t*-test
   E. Chi-square test of homogeneity

AP Questions

11. An advertising agency in a large city is conducting a survey of adults to investigate whether there is an association between highest level of educational achievement and primary source for news. The company takes a random sample of 2,500 adults in the city. The results are shown in the table below.

	HIGHEST LEVEL OF EDUCATIONAL ACHIEVEMENT			
Primary Source for News	Not High School Graduate	High School Graduate But Not College Graduate	College Graduate	Total
Newspapers	49	205	188	442
Local television	90	170	75	335
Cable television	113	496	147	756
Internet	41	401	245	687
None	77	165	38	280
Total	370	1,437	693	2,500

(a) If an adult is to be selected at random from this sample, what is the probability that the selected adult is a college graduate or obtains news primarily from the internet?
(b) If an adult who is a college graduate is to be selected at random from this sample, what is the probability that the selected adult obtains news primarily from the internet?
(c) When selecting an adult at random from the sample of 2,500 adults, are the events "is a college graduate" and "obtains news primarily from the internet" independent? Justify your answer.
(d) The company wants to conduct a statistical test to investigate whether there is an association between educational achievement and primary source for news for adults in the city. What is the name of the statistical test that should be used?
What are the appropriate degrees of freedom for this test?

12. Hurricane damage amounts, in millions of dollars per acre, were estimated from insurance records for major hurricanes for the past three decades. A stratified random sample of five locations (based on categories of distance from the coast) was selected from each of three coastal regions in the southeastern United States. The three regions were Gulf Coast (Alabama, Louisiana, Mississippi), Florida, and Lower Atlantic (Georgia, South Carolina, North Carolina). Damage amounts in millions of dollars per acre, adjusted for inflation, are shown in the table below.

HURRICANE DAMAGE AMOUNTS IN MILLIONS OF DOLLARS PER ACRE

	Distance from Coast				
	< 1 mile	1 to 2 miles	2 to 5 miles	5 to 10 miles	10 to 20 miles
Gulf Coast	24.7	21.0	12.0	7.3	1.7
Florida	35.1	31.7	20.7	6.4	3.0
Lower Atlantic	21.8	15.7	12.6	1.2	0.3

(a) Sketch a graphical display that compares the hurricane damage amounts per acre for the three different coastal regions (Gulf Coast, Florida, and Lower Atlantic) and that also shows how the damage amounts vary with distance from the coast.
(b) Describe differences and similarities in the hurricane damage amounts among the three regions.

Because the distributions of hurricane damage amounts are often skewed, statisticians frequently use rank values to analyze such data.
(c) In the table below, the hurricane damage amounts have been replaced by the ranks 1, 2, or 3. For each of the distance categories, the highest damage amount is assigned a rank of 1 and the lowest damage amount is assigned a rank of 3. Determine the missing ranks for the 10-to-20-miles distance category and calculate the average rank for each of the three regions. Place the values in the table below.

ASSIGNED RANKS WITHIN DISTANCE CATEGORIES

	Distance from Coast					
	< 1 mile	1 to 2 miles	2 to 5 miles	5 to 10 miles	10 to 20 miles	Average Rank
Gulf Coast	2	2	3	1		
Florida	1	1	1	2		
Lower Atlantic	3	3	2	3		

(d) Consider testing the following hypotheses.
H_0: There is no difference in the distributions of hurricane damage amounts among the three regions.
H_a: There is a difference in the distributions of hurricane damage amounts among the three regions.

If there is no difference in the distribution of hurricane damage amounts among the three regions (Gulf Coast, Florida, and Lower Atlantic), the expected value of the average rank for each of the three regions is 2. Therefore, the following test statistic can be used to evaluate the hypotheses above:

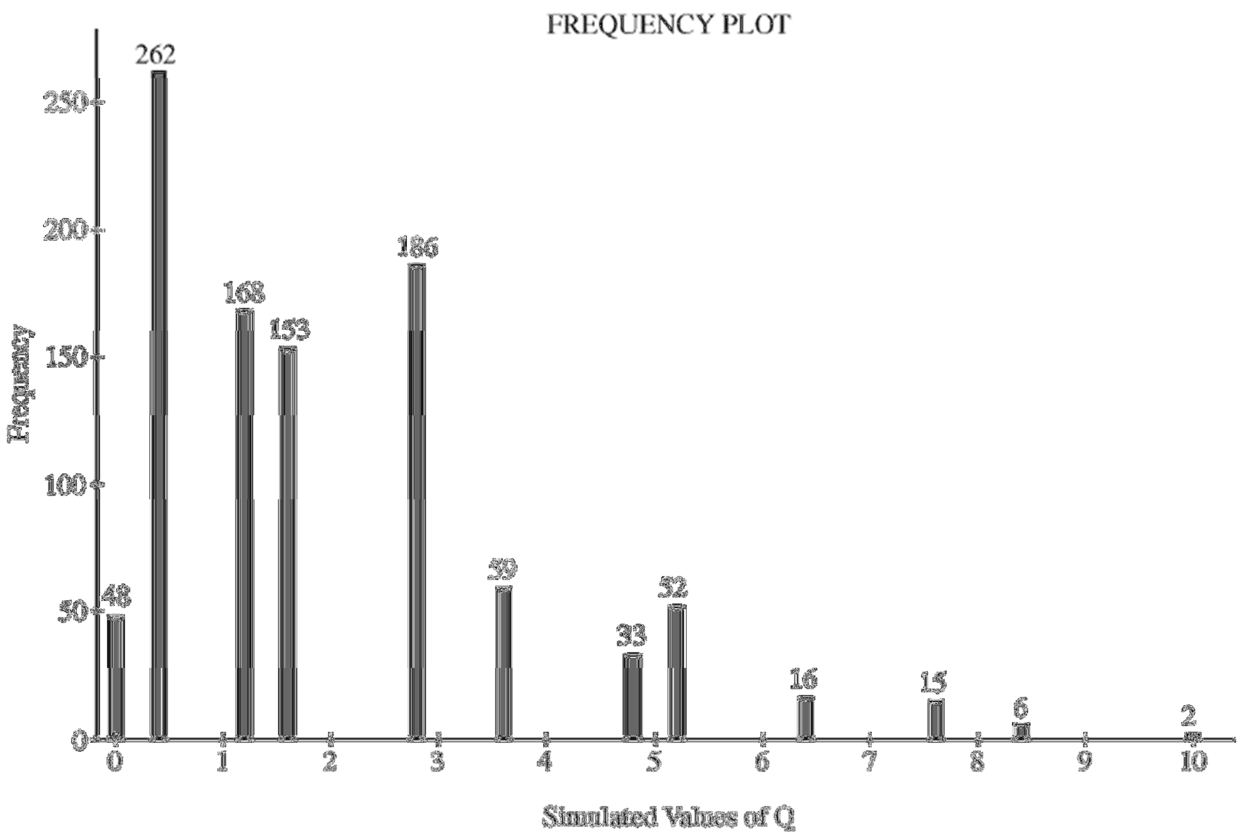
$$Q = 5[(\bar{R}_G - 2)^2 + (\bar{R}_F - 2)^2 + (\bar{R}_A - 2)^2]$$

where $\bar{R}_G$ is the average rank over the five distance categories for the Gulf Coast (and $\bar{R}_F$ and $\bar{R}_A$ are similarly defined for the Florida and Lower Atlantic coastal regions).
Calculate the value of the test statistic Q using the average ranks you obtained in part (c).

(e) One thousand simulated values of this test statistic, Q, were calculated, assuming no difference in the distribution of hurricane damage amounts among the three coastal regions. The results are shown in the table below. These data are also shown in the frequency plot where the heights of the lines represent the frequency of occurrence of simulated value of Q.

Frequency Table for Simulated Values of Q

Q	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0.0	48	48	4.80	4.80
0.4	262	310	26.20	31.00
1.2	168	478	16.80	47.80
1.6	153	631	15.30	63.10
2.8	186	817	18.60	81.70
3.6	59	876	5.90	87.60
4.8	33	909	3.30	90.90
5.2	52	961	5.20	96.10
6.4	16	977	1.60	97.70
7.6	15	992	1.50	99.20
8.4	6	998	0.60	99.80
10.0	2	1000	0.20	100.00



Use these simulated values and the test statistic you calculated in part (d) to determine if the observed data provide evidence of a significant difference in the distributions of hurricane damage amounts among the three coastal regions. Explain.