

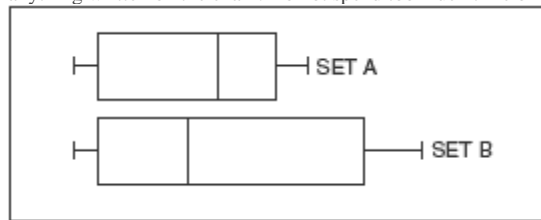
## SECTION I

Time: 1 hour and 30 minutes

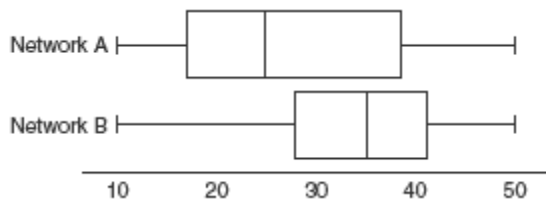
Number of questions: 40

Percentage of total grade: 50

**Directions:** Solve each of the following problems. Decide which is the best of the choices given and answer in the appropriate place on the answer sheet. No credit will be given for anything written on the exam. Do not spend too much time on any one problem.



1. Given the two boxplots above, which of the following statements is (are) true?
  - a. The boxplot for Set B has more terms above its median than the boxplot for Set A.
  - b. The boxplot for Set B has a larger IQR than the boxplot for Set A.
  - c. The median for Set A is larger than the median for Set B.
  - d. I only
  - e. II only
  - f. III only
  - g. I and II only
  - h. II and III only
2. An advertiser is trying to decide which television station in town to use for his product. He gathers the ratings of all prime time shows on each network and constructs a boxplot of each. There are the same number of ratings for each network. The results are as follows:



Based on these boxplots, which of the following is a correct conclusion about the relative ratings for the two networks?

- a. The median rating for Network A is greater than for Network B.
- b. The range for Network A is greater than for Network B.
- c. The interquartile ranges for the two networks are the same.
- d. The median rating for Network B is higher than for Network A.
- e. There are more ratings greater than 28 for Network A than for Network B.

3. A statistics class wanted to construct a 90% confidence interval for the difference in the number of advanced degrees held by male and female faculty members at their school. They collected degree data from all the male and female faculty members and then used these data to construct the desired 90% confidence interval. Is this an appropriate way to construct a confidence interval?
- . No, because we don't know that the distributions involved are approximately normal.
  - a. Yes, but only if the number of men and the number of women are equal because our calculations will be based on difference scores.
  - b. Yes, but only if the distribution of difference scores has no outliers or extreme skewness.
  - c. No, because all the data were available, there is no need to construct a confidence interval for the true difference between the number of degrees.
  - d. No, confidence intervals can only be constructed on independent samples, not on paired differences.
4. You are interested in determining which of two brands of tires (call them Brand G and Brand F) last longer under differing conditions of use. Fifty Toyota Camrys are fitted with Brand G tires and 50 Honda Accords are fitted with Brand F tires. Each tire is driven 20,000 miles, and tread wear is measured for each tire, and the average tread wear for the two brands is compared. What is wrong with this experimental design?
- . The type of car is a confounding variable.
  - a. Average tread wear is not a proper measure for comparison.
  - b. The experiment should have been conducted on more than two brands of cars.
  - c. Not enough of each type of tire was used in the study.
  - d. Nothing is wrong with this design—it should work quite well to compare the two brands of tires.
5. The blood types of 468 people residing in the United States (all of whom were Asian, African-American, Arab, or White) were collected in a study to see if their blood type distribution is related to race. The following results were obtained:

Type Ethnicity	O	A	B	AB
Asian	48	33	33	6
African-American	45	24	18	3
Arab	36	33	30	6
White	69	60	18	6

- From these data, a  $\chi^2$  value of 19.59 (df = 9) was computed. At the 5% level of significance, do these data indicate that Ethnicity and Blood Type are related?
- . Yes, because the  $P$ -value of the test is greater than 0.05.
  - a. Yes, because the  $P$ -value of the test is less than 0.05.
  - b. No, because the  $P$ -value of the test is greater than 0.05.
  - c. No, because the  $P$ -value of the test is less than 0.05.
  - d.  $\chi^2$  should not be used in this situation since more than 20% of the expected values are less than 5.
6. Most college-bound students take either the SAT (Scholastic Assessment Test) or the ACT (which originally stood for American College Testing). Scores on both the ACT and the SAT are approximately normally distributed. ACT scores have a mean of about 21 with a standard deviation of about 5. SAT scores have a mean of about 508 with a standard deviation of about 110. Nicole takes the ACT and gets a score of 24. Luis takes the SAT. What score would Luis have to have on the SAT to have the same standardized score (z-score) as Nicole's standardized score on the ACT?
- . 548
  - a. 574
  - b. 560

c. 583

d. 588

7. A researcher conducts a study of the effectiveness of a relaxation technique designed to improve the length of time a SCUBA diver can stay at a depth of 60 feet with a 80 cu. ft. tank of compressed air. The average bottom time for a group of divers before implementation of the program was 48 minutes and the average bottom time after implementation of the program was 54 minutes with a  $P$ -value of 0.024. Which of the following is the best interpretation of this finding?

- . There is a 2.4% chance that the new technique is effective at increasing bottom time.
- a. If the new technique was not effective, there is only a 2.4% chance of getting 54 minutes or more by chance alone.
- b. 97.6% of the divers in the study increased their bottom times.
- c. We can be 97.6% confident that the new technique is effective at increasing bottom time.
- d. The new technique does not appear to be effective at increasing bottom time.

8. Does ultraviolet radiation affect the birth rate of frogs? A study in the *Tampa Tribune* reported that while 34 of 70 sun-shaded (from ultraviolet radiation) eggs hatched, only 31 of 80 unshaded eggs hatched. Which of the following would give a 99% confidence interval for the true difference between the proportions of shaded and unshaded eggs that hatched?

a.  $\left(\frac{34}{70} - \frac{31}{80}\right) \pm 1.96 \sqrt{\frac{\frac{34}{70}\left(1 - \frac{34}{70}\right)}{70} + \frac{\frac{31}{80}\left(1 - \frac{31}{80}\right)}{80}}$

b.  $\left(\frac{34}{70} - \frac{31}{80}\right) \pm 2.576 \sqrt{\left(\frac{65}{150}\right)\left(\frac{85}{150}\right)\left(\frac{1}{70} + \frac{1}{80}\right)}$

c.  $\left(\frac{34}{70} - \frac{31}{80}\right) \pm 1.96 \sqrt{\left(\frac{65}{150}\right)\left(\frac{85}{150}\right)\left(\frac{1}{70} + \frac{1}{80}\right)}$

d.  $\left(\frac{34}{70} - \frac{31}{80}\right) \pm 2.576 \sqrt{\left(\frac{65}{150}\right)\left(\frac{85}{150}\right)\left(\frac{34}{70} + \frac{31}{80}\right)}$

e.  $\left(\frac{34}{70} - \frac{31}{80}\right) \pm 2.576 \sqrt{\frac{\frac{34}{70}\left(1 - \frac{34}{70}\right)}{70} + \frac{\frac{31}{80}\left(1 - \frac{31}{80}\right)}{80}}$

9. At Midtown University, the average weight of freshmen boys is 170 lbs with a standard deviation of 9 lbs. The average weight of freshmen girls is 115 lbs with a standard deviation of 6 lbs. A new distribution is to be formed of the values obtained when the weights of the girls and the boys are added together. What are the mean and standard deviation of this new distribution? Assume that the weights of freshman boys and freshman girls are independent.

. 285, 15

a. 285, 117

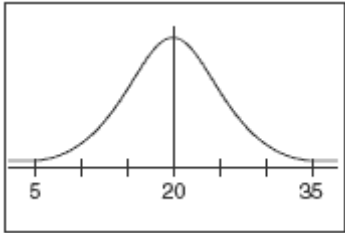
b. 55, 10.8

c. 285, 10.8

d. The mean is 285 but, under the conditions stated in the problem, you cannot determine the standard deviation.

10. A random sample of 875 deaths in the United States in the year 2000 showed a mean life span of 75.1 years with a sample standard deviation of 16 years. These data were used to generate a 95% confidence interval for the true mean lifespan in the United States. The interval constructed was (74.0, 76.2). Which of the following statements is correct?

- . There is a 95% chance that the average lifespan in the United States is between 74 and 76.2 years.
- a. 95% percent of the time, a person in the United States will live between 74 years and 76.2 years.
- b. 95% of the time, on average, intervals produced in this manner will contain the true mean lifespan.
- c. On average, 95% of people live less than 76.2 years.
- d. The probability is 0.95 that this interval contains the true mean lifespan in the United States.



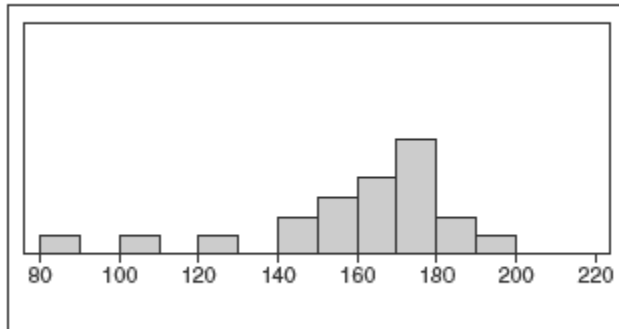
11.

Which of the following is the best estimate of the standard deviation for the approximately normal distribution pictured?

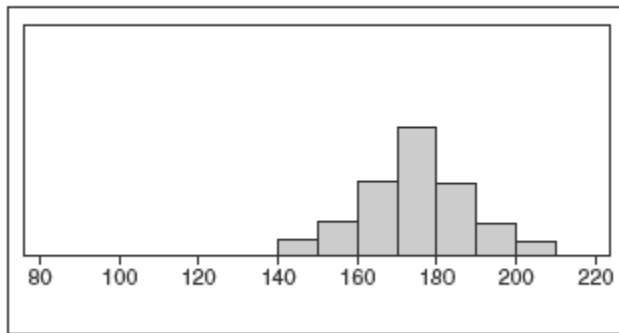
- . 10
- a. 30
- b. 5
- c. 9
- d. 15

12. The following histograms compare two datasets (A and B):

**Sample A**



**Sample B**

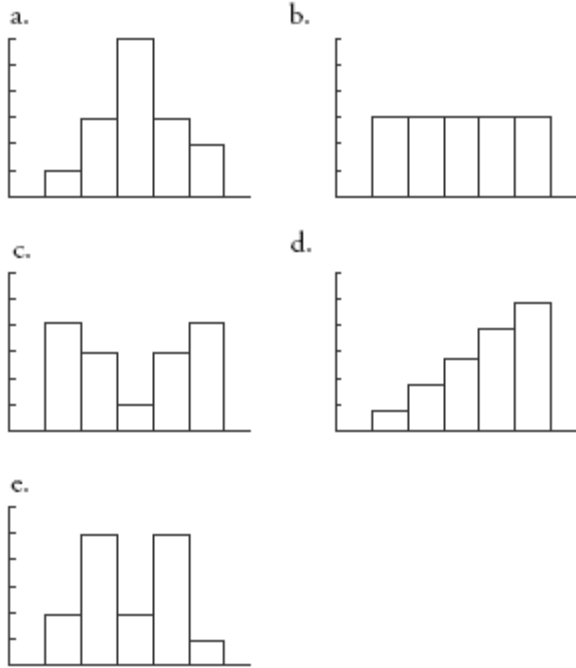


Which of the following statements is true?

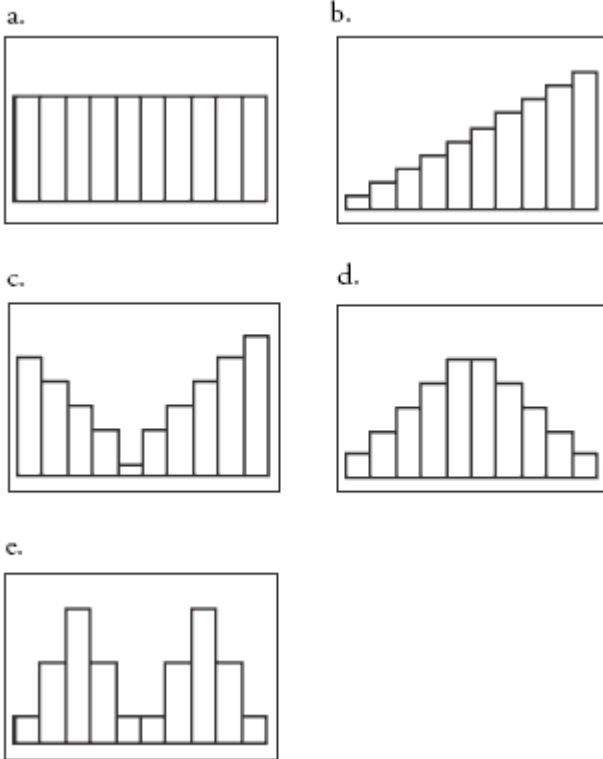
- . Sample A has more values than Sample B.
  - a. The mean of Sample A is greater than the mean of Sample B.
  - b. The mean of Sample B is greater than the mean of Sample A.
  - c. The median of Sample A is greater than the median of Sample B.
  - d. Both graphs are symmetric about their mean.
13. Sometimes, the order in which a question is asked makes a difference in how it is answered. For example, if you ask people if they prefer chocolate or strawberry ice cream, you might get different answers than if you asked them if they prefer strawberry or chocolate ice cream. Seventy-five randomly selected people were asked, "Do you prefer chocolate or strawberry?" and 75 different randomly selected people were asked, "Do you prefer strawberry or chocolate?" The results are given in the following table.

	<b>RESPONSE: PREFER CHOCOLATE</b>	<b>RESPONSE: PREFER STRAWBERRY</b>
<b>Question:</b> Do you prefer chocolate or strawberry?	42	33
<b>Question:</b> Do you prefer strawberry or chocolate?	34	41

- A two-proportion z-test was performed on these data to see if the order of the question made a difference. What is the  $P$ -value of the test (hint: you need to think about whether this is a one-sided or a two-sided test)?
- . 0.453
  - a. 0.096
  - b. 0.56
  - c. 0.055
  - d. 0.19
14. Young Aheart is interested in the extent to which teenagers might favor a dress code for high school students. He has access to a list of 55,000 teens in a large urban district. He draws a random sample of 40 students from this list and records the count of those who say they favor a dress code. He then repeats this process 24 more times. What kind of distribution has he simulated?
- . The sampling distribution of a sample proportion with  $n = 25$
  - a. The sampling distribution of a sample mean with  $n = 40$
  - b. The binomial distribution with  $n = 25$
  - c. The sampling distribution of a sample proportion with  $n = 40$
  - d. The geometric distribution
15. Each of the histograms below is of 15 integers from 1 through 5. The horizontal and vertical scales are the same for each graph. Which graph has the *smallest* standard deviation?



16. Which of the five histograms pictured below has the *smallest* standard deviation?



17. Alligators captured in Florida are found to have a mean length of 2 meters and a standard deviation of 0.35 meters. The lengths of alligators are believed to be approximately normally distributed. What is the approximate length of an alligator at the 67th percentile of alligator lengths?

- . 2.01 meters.

- a. 2.44 meters.
- b. 2.21 meters.
- c. 2.15 meters.
- d. 2.09 meters.

$$r = \frac{1}{n-1} \sum \left( \frac{x-\bar{x}}{s_x} \right) \left( \frac{y-\bar{y}}{s_y} \right) ?$$

18. Which of the following is not a property of the correlation coefficient
- .  $r$  is not a function of the units used for the variables.
  - a.  $r$  can be calculated from either categorical or numerical variables.
  - b.  $r$  is not affected by which variable is called  $x$  and which variable is called  $y$ .
  - c.  $|r| \geq 1$ .
  - d.  $r$  is positive when the slope of the regression line is positive and negative when the slope of the regression line is negative.
19. A study published in the *Journal of the National Cancer Institute* (Feb. 16, 2000) looked at the association between cigar smoking and death from cancer. The data reported were as follows:

Death from Cancer			
	Yes	No	Total
Never Smoked	782	120,747	121,529
Former Smoker	91	7,757	7,848
Current Smoker	141	7,725	7,866
Total	1,014	136,229	137,243

Which of the following statements is true?

- . A former smoker is more likely to have died from cancer than a person who has never smoked.
  - a. Former smokers and current smokers are equally likely to have died from cancer.
  - b. The events "Current Smoker Dies from Cancer" and "Died from Cancer" are independent.
  - c. It is more likely that a person who is a current smoker dies from cancer than a person has never smoked and dies from cancer.
  - d. Among those whose death was not from cancer, the proportion of current smokers is higher than the proportion of former smokers.
20. Which of the following is a reason for choosing a z-procedure rather than a t-procedure when making an inference about the mean of a population?
- . The standard deviation of the population is unknown.
  - a. The sample was a simple random sample.
  - b. The sample size is greater than 40.
  - c. The shape of the population from which the sample is drawn is approximately normal.
  - d. The population standard deviation is known.
21. You play a game that involves rolling a die. You either win or lose \$1 depending on what number comes up on the die. If the number is even, you lose \$1, and if it is odd, you win \$1. However, the die is weighted and has the following probability distribution for the various faces:

Face	1	2	3	4	5	6
Win ( $x$ )	+1	-1	+1	-1	+1	-1
$p(x)$	0.15	0.20	0.20	0.25	0.1	0.1

Given that you win rather than lose, what is the probability that you rolled a "5"?

- . 0.50
  - a. 0.10
  - b. 0.45
  - c. 0.22
  - d. 0.55
22. A psychiatrist is studying the effects of regular exercise on stress reduction. She identifies 40 people who exercise regularly and 40 who do not. Each of the 80 people is given a questionnaire designed to determine stress levels. None of the 80 people who participated in the study knew that they were part of a study. Which of the following statements is true?
- . This is an observational study.
  - a. This is a randomized comparative experiment.
  - b. This is a double-blind study.
  - c. This is a matched-pairs design.
  - d. This is an experiment in which exercise level is a blocking variable.
23. It is the morning of the day that Willie and Baxter have planned their long-anticipated picnic. Willie reads, with some distress, that there is a 65% probability of rain in their area today.
- Which of the following best describes the most likely way that probability was arrived at?
- . It rains 65% of the time on this date each year.
  - a. Historically, in the United States, it has rained 65% of the time on days with similar meteorological conditions as today.
  - b. Historically, it rains 65% of the days during this month.
  - c. Historically, in this area, it has rained 65% of the time on days with similar meteorological conditions as today.
  - d. This is the result of a simulation conducted by the weather bureaus.
24. In order to meet air pollution standards, the mean emission level for engines of a certain type must be less than 20 parts per million (ppm) of carbon. A study is to be done to determine if the engines from a particular company meet the standard. Which of the following represents the correct null and alternative hypotheses for this study? Let  $\mu$  = the mean parts/million of carbon emitted for these cars.
- .  $H_0 : \mu = 20; H_A : \mu > 20$
  - a.  $H_0 : \mu \geq 20; H_A : \mu < 20$
  - b.  $H_0 : \mu > 20; H_A : \mu < 20$
  - c.  $H_0 : \mu = 20; H_A : \mu \geq 20$
  - d.  $H_0 : \mu \geq 20; H_A : \mu > 20$
25. Given  $P(A) = 0.60$ ,  $P(B) = 0.30$ , and  $P(A|B) = 0.50$ . Find  $P(A \cup B)$ .
- . 0.90
  - a. 0.18
  - b. 0.40
  - c. 0.72



d. 0.75

Use the following information to answer questions 26–27:

Baxter is a 60% free-throw shooter who gets fouled during a game and gets to shoot what is called a "one-and-one" (that is, he gets to take a second shot—a bonus—if and only if he makes his first shot; each free throw, if made, is worth one point). Baxter can make 0 points (because he misses his first shot), 1 point (he makes the first shot, but misses the bonus), or 2 points (he makes his first shot and the bonus).

26. Assuming that each shot is independent, how many points is Baxter *most* likely to make in a one-and-one situation?

. 2

a. 1

b. 0

c. 0.96

d. None of these is correct.

27. Assuming that each shot is independent, how many points will Baxter make *on average* in a one-and-one situation?

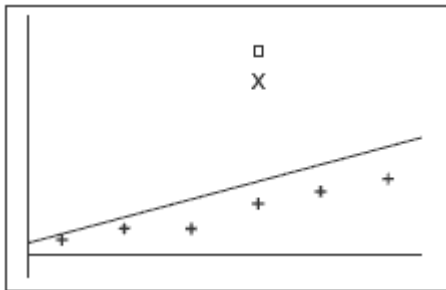
. 2

a. 0.96

b. 0

c. 1

d. 0.36



28.

For the graph given above, which of the following statements is (are) true?

. The point marked with the "X" is better described as an outlier than as an influential point.

a. Removing the point "X" would cause the correlation to increase.

b. Removing the point "X" would have a significant effect on the slope of the regression line.

c. I and II only

d. I only

e. II only

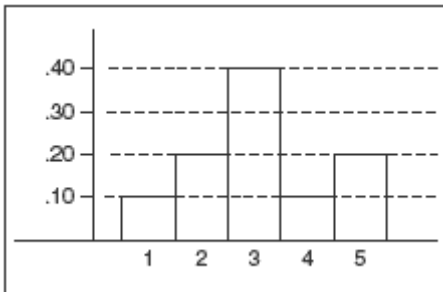
f. II and III only

g. I, II, and III

29. A two-proportion large-sample confidence interval is to be constructed. Which of the following is *not* usually considered necessary to construct such an interval?

. The two samples should be SRSs from their respective populations.

- a. The two samples are independent.
  - b. The populations from which the samples are drawn should each be approximately normal.
  - c.  $n_1 \hat{p}_1 \geq 5, n_1(1 - \hat{p}_1) \geq 5, n_2 \hat{p}_2 \geq 5, n_2(1 - \hat{p}_2) \geq 5$ .
  - d. The critical value is always  $z$  rather than  $t$ .
30. Results of an experiment or survey are said to be *biased* if
- . Subjects are not assigned randomly to treatment and control groups.
  - a. Some outcomes are systematically favored over others.
  - b. There was no control group.
  - c. A double-blind procedure was not used.
  - d. The sample size was too small to control for sampling variability.



31. Given the probability histogram pictured for a discrete random variable  $X$ , what is  $\mu_x$  ?
- . 3.0
  - a. 0.25
  - b. 2.5
  - c. 3.1
  - d. 2.8
32. A fair die is to be rolled 8 times. What is the probability of getting at least one 4?
- a.  $\frac{1}{6}$
  - b.  $\binom{8}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^7$
  - c.  $1 - \binom{8}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^7$
  - d.  $\binom{8}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^6 + \binom{8}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^5 + \dots + \binom{8}{8} \left(\frac{1}{6}\right)^8 \left(\frac{5}{6}\right)^0$
  - e.  $1 - \binom{8}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^8$

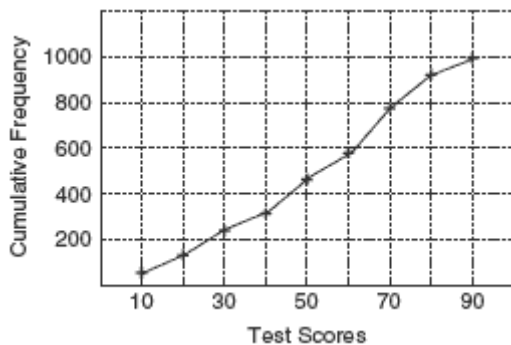
33. Which of the following statements is not true about the power of a statistical test?
- . You can increase the power of a test by increasing the significance level.

- a. When  $H_0$  is false, power =  $1 - \alpha$ .
  - b. The power of a test equals the probability of rejecting the null hypothesis.
  - c. You can increase the power of a test by increasing the sample size.
  - d. The power of a test is a function of the true value of the parameter being tested.
34. 40% of the staff in a local school district have a master's degree. One of the schools in the district has only 4 teachers out of 15 with a master's degree. You are asked to design a simulation to determine the probability of getting this few teachers with master's degrees in a group this size. Which of the following assignments of the digits 0 through 9 would be appropriate for modeling this situation?
- . Assign "0,1,2" as having a master's degree and "4,5,6,7,8,9" as not having a degree.
  - a. Assign "1,2,3,4,5" as having a master's degree and "0,6,7,8,9" as not having a degree.
  - b. Assign "0,1" as having a master's degree and "2,3,4,5,6,7,8,9" as not having a degree.
  - c. Assign "0,1,2,3" as having a master's degree and "4,5,6,7,8,9" as not having a degree.
  - d. Assign "7,8,9" as having a master's degree and "0,1,2,3,4,5,6," as not having a degree.
35. Which of the following statements is (are) true about the  $t$ -distribution?
- . Its mean, median, and mode are all equal.
  - a. The  $t$ -distribution is more spread out than the  $z$ -distribution.
  - b. The greater the number of degrees of freedom, the greater the variance of a  $t$ -distribution.
  - c. I only
  - d. II only
  - e. III only
  - f. I and II only
  - g. I and III only
36. A study showed that persons who ate two carrots a day have significantly better eyesight than those who eat fewer than one carrot a week. Which of the following statements is (are) correct?
- . This study provides evidence that eating carrots contributes to better eyesight.
  - a. The general health consciousness of people who eat carrots could be a confounding variable.
  - b. This is an observational study and not an experiment.
  - c. I only
  - d. III only
  - e. I and II only
  - f. II and III only
  - g. I, II, and III
37. You are designing a study to determine which of three brands of golf ball will travel the greatest distance. You intend to use only adult male golfers. There is evidence to indicate that the temperature at the time of the test affects the distance traveled. There is no evidence that the size of the golfer is related to the distance traveled (distance seems to have more to do with technique than bulk). This experiment would best be done
- . by blocking on type of golf ball.
  - a. by blocking on size of the golfer.
  - b. by blocking on size of the golfer and temperature.
  - c. without blocking.
  - d. by blocking on temperature.

38. Given the cumulative frequency table shown below, what are the mean and median of the distribution?

VALUE	CUMULATIVE FREQUENCY
2	0.15
3	0.25
5	0.45
7	0.95
10	1.00

- . Mean = 5.6, median = 7
- a. Mean = 5.6, median = 5
- b. Mean = 5.4, median = 7
- c. Mean = 5.4, median = 5
- d. Mean = 4.8, median = 6
39. A spelling test was given to 1000 elementary students in a large urban school district. The graph below is a cumulative frequency graph of the results. Which of the following is closest to the five-number summary (minimum, first quartile, median, third quartile, maximum) for the distribution of spelling scores?



- . {10, 30, 50, 70, 90}
- a. {10, 20, 50, 80, 90}
- b. {0, 30, 50, 70, 100}
- c. {20, 40, 60, 80, 100}
- d. There is not enough information contained in the graph to determine the five-number summary.
- A well-conducted poll showed that 46% of a sample of 1500 potential voters intended to vote for Geoffrey Sleazy for governor. The poll had a reported margin of error of 3%. Which of the following best describes what is meant by "margin of error of 3%"?
    - f. The probability is 0.97 that between 43% and 49% will vote for candidate Sleazy.
    - g. Ninety-seven percent of the time, between 43% and 49% would vote for candidate Sleazy.
    - h. Between 43% and 49% of voters will vote for Sleazy.
    - i. Three percent of those interviewed refused to answer the question.
    - j. The proportion of voters who will vote for Sleazy is likely to be between 43% and 49%.

## SECTION II

Time: 1 hour and 30 minutes

Number of problems: 6

Percentage of total grade: 50

### General Instructions

There are two parts to this section of the examination. Part A consists of five equally weighted problems that represent 75% of the total weight of this section. Spend about 65 minutes on this part of the exam. Part B consists of one longer problem that represents 25% of the total weight of this section. Spend about 25 minutes on this part of the exam. You are not necessarily expected to complete all parts of every question. Statistical tables and formulas are provided.

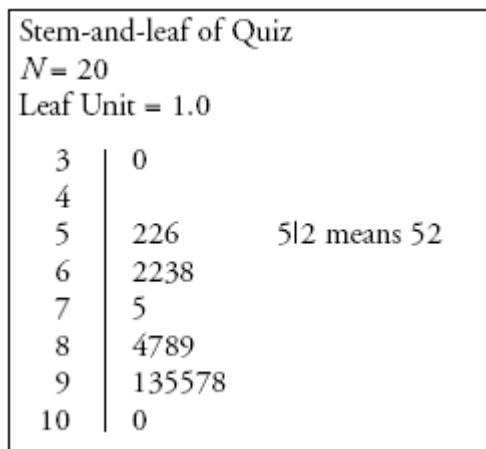
- Be sure to write clearly and legibly. If you make an error, you may save time by crossing it out rather than trying to erase it. Erased or crossed-out work will not be graded.
- Show all your work. Indicate clearly the methods you use because you will be graded on the correctness of your methods as well as the accuracy of your final answers. Correct answers without support work may not receive credit.

### Statistics, Section II, Part A, Questions 1–5

Spend about 65 minutes on this part of the exam; percentage of Section II grade: 75.

**Directions:** Show all your work. Indicate clearly the methods you use because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

1. The following stemplot gives the scores of 20 statistics students on the first quiz of the quarter.



- a. What is the median score on this quiz for the 20 students?
  - b. Draw a boxplot of the data.
  - c. What is the lowest score a student could get on this quiz and still not be an outlier? Explain.
2. The 1970 draft lottery involved matching birthdates with a number from 1 to 366. The lower the number, the more likely the individual with the matching birthday was to be drafted to fight in Vietnam. The average selection numbers by month are given in the following table.

MONTH	AVERAGE NUMBER
January	201.2
February	203.0
March	225.8
April	203.7
May	208.0
June	195.7
July	181.5
August	173.5
September	157.3
October	182.5
November	148.7
December	121.5

The following is part of the computer output for the least-squares regression line for predicting draft number from birth month (January = 1, February = 2, etc.).

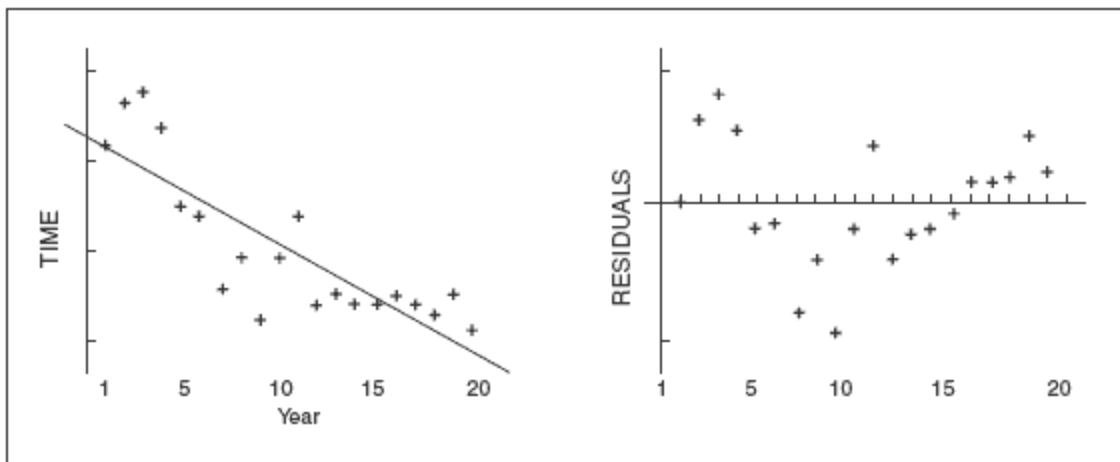
Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	229.402	9.466	24.23	.000
Month	-7.057	1.286	-5.49	.000

- a. What is the equation of the least-squares regression line for predicting average draft number from birth month?
  - b. Interpret the slope of the regression line in the context of the problem.
  - c. Does the computer analysis indicate that there is a useful (predictive) relationship between birth month and average draft number? Explain.
  - d. Construct a 95% confidence interval for the true slope of the regression line and interpret the interval in the context of the problem.
3. A sleep researcher wants to know if people get a better quality sleep on expensive mattresses than on economy mattresses. The researcher obtains 125 volunteers for the study. Eighty of the volunteers purchased their mattresses at EconoSleep, a heavily advertised store that caters to working families and the other 45 bought their mattresses at Night of Luxury, an upscale sleep store. The 125 volunteers were given a questionnaire to determine the quality of sleep on each mattress. The responses were then compared and a report was prepared comparing the two type of mattresses.
- a. Is this an experiment or an observational study? Explain.
  - b. Explain what is meant by a confounding variable in the context of this study. Give an example of a possible confounding variable.
  - c. Suppose the report concluded that people who bought their mattresses from Night of Luxury slept better than those who bought their mattresses from EconoSleep. Could you conclude that the difference was attributable to the quality of the mattresses in the study? Explain.
4. The National Collegiate Athletic Association (NCAA) issues reports on the graduation rates at its member schools. The following table gives the graduation rates of female students and female basketball players at several randomly selected colleges.

SCHOOL	GRADUATION RATE FOR ALL FEMALE STUDENTS	GRADUATION RATES FOR ALL FEMALE BASKETBALL PLAYERS
A	59	67
B	54	75
C	39	52
D	71	55
E	69	40
F	74	55
G	64	67
H	70	40
I	40	50
J	85	95

Do these data provide evidence that there is a difference between the mean graduation rates for all female students and the mean graduation rates for all female basketball players? Give good statistical evidence for your response.

5. The graphs below give the times of the winner of the women's race at the Boston Marathon from 1975 to 1994. The graph on the left is the scatterplot of the year (numbered from 1 so that 1975 = 1, 1976 = 2, etc.) versus the times (which range from 2 hours 48 minutes down to 2 hours, 22 minutes). The graph on the right is a plot of the residuals versus the year. The equation of the regression line is  $\text{Time} = 163.1 - 1.23(\text{Year})$ , where the year is the number of years after 1975.



- What would be the predicted time for the winner in 1999 (that would be year 25)?
- The winner in 1999 actually ran the race in 143 minutes. What is the residual for this prediction?
- Does a line appear to be a good model for the data? Explain.
- If your goal was to predict the time for the winner in 1999, suggest an improvement to the situation described above but using the same data, or part of the same data.

**Statistics, Section II, Part B, Question 6**

Spend about 25 minutes on this part of the exam; percentage of Section II grade: 25

**Directions:** Show all of your work. Indicate clearly the methods you use because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

6. The casino game of *Chuck-a-Luck* involves spinning a cage with three dice in it. You bet \$1 on a single number from 1–6 and you win \$1 for each face-up appearance of your number. Because it is a bet, you lose your dollar only if your number appears on none of the three dice (e.g., suppose you bet on "5" and it come up on two dice—you receive \$2 and get to keep your original bet). There are 216 ways for three dice to come up ( $6 \times 6 \times 6$ ). Let  $X$  = the amount won on each play of the game.
  - a. Construct the probability distribution for *Chuck-a-Luck*. (Hint: there are four values of  $X$ )
  - b. What are your expected mean and standard deviation for each \$1 that you bet?
  - c. Suppose you hang around the *Chuck-a-Luck* game for a couple of hours and play 100 times. What is the probability that you make a profit? Suppose that you were able to place 10,000 bets on Chuck-a-Luck. Now, what is the probability that you make a profit?

### Solutions to Practice Exam 2, Section I

1. The correct answer is (e). I is not correct because there is no way to tell from a boxplot how many terms are in a dataset. Thus, you can't compare the number above the median in A with the number above the median in B.
2. The correct answer is (d). (a) is not correct since the median for B is greater than for A. (b) is not correct since the ranges of the two networks appears to be the same. (c) is not correct since the IQR for Network A is greater than the IQR for Network B. (e) is not correct since about 75% of the ratings for Network B are greater than 28 but less than half of those for Network A are greater than 28.
3. The correct answer is (d). Sample data are used to generate confidence intervals for unknown population values. In this case, we have all of the population values so that we can compute the difference in the number of degrees between male and female faculty members. It make no sense to construct a confidence interval if you have all the data about the population.
4. The correct answer is (a). It's possible that the type of car the tire is on influences the tread wear and that hasn't been controlled for in the study. A better design would be to block by type of car. Then 25 of each model would be randomly assigned tire F and 25 would randomly be assigned tire G. Then comparisons would be made within each block.
5. The correct answer is (b). Using Table C for  $df = 9$ , we find that 19.59 is between 19.02 and 19.68. Looking to the top of the columns in which those numbers appear, we see that  $0.02 < P < 0.025$ . Hence, the  $P$ -value is less than 0.05, which provides statistical evidence that Ethnicity and Blood Type are related at the 5% level of significance. On the TI-83/84, we have  $P\text{-value} = X^2 \text{cdf}(19.59, 1000, 9) = 0.0206$ . Note that two of the expected values (for African-American and Arab of Type AB) are less than 5. However, this is less than 20% of all expected values so that we are justified in proceeding with the analysis.

6. The correct answer is (b). Nicole's  $z$ -score on the ACT is  $z = \frac{24 - 21}{5} = 0.6$ . Let Luis's score on the SAT be  $x$ . Then, Luis's  $z$ -score on the SAT is  $z = \frac{x - 508}{110}$ . Thus,  $\frac{x - 508}{110} = 0.6 \Rightarrow x = 574$ .

7. The correct answer is (b). The  $P$ -value of a finding is the probability of getting a finding as or more extreme than the one obtained by chance alone if the null hypothesis is true. (a), (c), and (d) are simply misstatements of what  $P$ -value means. The  $P$ -value is relatively low (0.024), which means that a conclusion like that in part (e) would be arguable. Most would agree that such a low  $P$ -value *does* provide evidence that the technique is effective.
8. The correct answer is (e). 1.96 would be the correct value of  $\alpha$  if the question asked for a 95% confidence interval (this rules out (a) and (c)), (b), (c), and (d)) all use a pooled estimate of  $p$ , which is correct (at least in (b) and (c)) for a hypothesis test but not for a confidence interval.
9. The correct answer is (d).

$$\mu_{x+y} = \mu_x + \mu_y = 170 + 115 = 285;$$

since  $X$  and  $Y$  are independent,



$$\sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2} = \sqrt{9^2 + 6^2} = 10.8.$$

10. The correct answer is (c). This is exactly what is meant by a "95% confidence interval." The other answers all involve some sort of probability statement about the constructed interval. Once the interval is constructed, the true mean is either in the interval (probability = 1) or it isn't (probability = 0). Note the difference between the statements, "The probability is 0.95 that *this* interval contains the true mean" and "The probability is 0.95 that an interval *constructed in this manner* will contain the true mean."
11. The correct answer is (c). It appears from the graph that most of the data will lie in the interval (5, 35), which means the range of values is about 30. In a normal curve, almost all of the data are within three standard deviations of the mean, or a range of six standard deviations total. Thus,  $30/6 = 5$  is the best estimate of the standard deviation.
12. The correct answer is (c). (a) is not correct since there is no way to tell from a boxplot how many values are in a data set. It is evident from the graphs and the negative skewness of Sample A, that the mean of Sample B will be greater than the mean of Sample A since the mean will be pulled in the direction of skewness. Hence (b) is incorrect and (c) is correct. (d) is incorrect—the median for Sample A is to the left of the median for Sample B. (e) is incorrect since only Sample B is symmetric about its mean.
13. The correct answer is (e).

$$H_0: p_1 - p_2 = 0, H_A: p_1 - p_2 \neq 0$$

$$\hat{p}_1 = \frac{42}{75} = 0.56, \quad \hat{p}_2 = \frac{34}{75} = 0.45, \quad \hat{p} = \frac{42 + 34}{75 + 75} = 0.51$$

$$z = \frac{0.56 - 0.45}{\sqrt{0.51(1 - 0.51)\left(\frac{1}{75} + \frac{1}{75}\right)}} = 1.31 \Rightarrow P\text{-value} =$$

$$2(1 - 0.9049) = 0.1902 \text{ (from Table A; on the TI-83/84, } 2 \times \text{normalcdf}(1.31, 100) = 0.1902).$$

(Note that the given answer is based on the accuracy provided by the TI-83/84 2-PropZ Test function. If you used only two-plate accuracy and plugged the numbers given in the formula for  $z$  into the calculator, you would get a  $z$ -score of 1.35 and a  $P$ -value of 0.18.)

14. The correct answer is (d). He is drawing multiple samples, each of size 40. For each sample, he is computing a sample proportion (the count of successes/40). He would have to take all possible samples of size 40 to actually have the sampling distribution.
15. The correct answer is (a). The graph with the *smallest* standard deviation is the graph that is most packed about its mean (varies least from the center). Note that (b) would be the graph with the *greatest* standard deviation.
16. The correct answer is (d). The more the terms pack around the center of the distribution, the smaller the spread  $\Rightarrow$  the smaller the standard deviation. (d) has more terms bunched in the center than the other graphs. The graph with the most variation appears to be (c).
17. The correct answer is (d). Let  $x$  be the unknown length of an alligator at the 67th percentile. Then  $z_x = 0.44$  (from Table A; using the TI-83/84,  $z_x = \text{invNorm}(0.67) = 0.4399$ ). Also,  $z_x = \frac{x - 2}{0.35}$ . Equating the two expressions for  $z$ , we find  $x = 2.15$  meters.

18. The correct answer is (b). The formula for  $r$  requires the use of numerical variables.

$$= \frac{91}{7848} = 0.012$$

19. The correct answer is (a).  $P(\text{former smoker dies from cancer}) = \frac{121,529}{782}$ .  $P(\text{a person who has never smoked dies from$

cancer) =  $\frac{121,529}{7,866} = 0.006$ . (b) is false since  $P(\text{former smoker dies from cancer}) = 0.012$  and  $P(\text{current smoker dies from$

cancer) =  $\frac{141}{7,866} = 0.018$

(c) is false since  $P(\text{current smoker dies from cancer}) = 0.028$  and  $P(\text{current smoker dies from cancer} |$

person died from cancer) =  $\frac{141}{1,014} = 0.139$

Since these probabilities are not the same, the events are not independent. (d) is false

$$= \frac{141}{137,243} = 0.001$$

since  $P(\text{current smoker dies and dies from cancer}) = \frac{782}{137,243} = 0.006$  and  $P(\text{a person who never smoked and dies from cancer}) = \frac{7,725}{136,229} = 0.0567$ . (e) is false since  $P(\text{current smoker} \mid \text{did not die from cancer}) = \frac{7,757}{136,229} = 0.0569$  (but: very close!).

20. The correct answer is (e). Use of a  $z$ -test assumes that the population standard deviation is known. Be sure you understand that choice (c) is not a reason for choosing  $z$  even though, with a sample that large, the  $P$ -value obtained will be very close. For large samples, the population standard deviation is a good estimate of the population standard deviation, but that doesn't make it a  $z$ -procedure—just a stronger justification for using  $t$ .
21. The correct answer is (d).

$$P(5 \mid \text{win}) = P(5 \mid 1 \text{ or } 3 \text{ or } 5) = \frac{0.1}{0.15 + 0.20 + 0.1} = \frac{10}{45} = 0.22.$$

22. The correct answer is (a). Note that this is not an experiment since no treatment is imposed.
23. The correct answer is (d). The probability of rain on a given day is computed as the relative frequency of days that it has actually rained in the past when conditions like those in effect today have been present.
24. The correct answer is (b). The hypothesis is one sided and the researcher is interested in knowing if cars produce *less* than 20 ppm. That eliminates (a), (d), and (e) (since the alternative points the wrong way). A null hypothesis must have an equal sign, which means (b) is correct. Note that the correct answer could equally well be stated  $H_0: \mu = 20$ ;  $H_A: \mu < 20$ .
25. The correct answer is (e).  $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.60 + 0.30 - P(A \cap B) = 0.90 - P(A \cap B)$ . Now,  $P(A \cap B) = P(B) \cdot P(A \mid B) = (0.30)(0.50) = 0.15$ . Thus,  $P(A \cup B) = 0.90 - P(A \cap B) = 0.90 - 0.15 = 0.75$ .
26. The correct answer is (c).  $P(0) = 0.4$ ;  $P(1) = (0.6)(0.4) = 0.24$ ;  $P(2) = (0.6)(0.6) = 0.36$ . So the *most* likely number of points scored is 0.
27. The correct answer is (b). From the previous problem, we have the following probability distribution for this situation.

<b>Points Made (<math>x</math>)</b>	0	1	2
<b><math>p(x)</math></b>	0.4	0.24	0.36

$$\mu_x = (0)(0.4) + (1)(0.24) + 2(0.36) = 0.96.$$

28. The correct answer is (a). The removal of point "X" would have a minimal effect on the slope of the regression line, but a major effect on the correlation coefficient and the intercept of the regression line.
29. The correct answer is (c). The normality condition is satisfied by the equations in answer (d). It is not necessary that the original populations be approximately normal.
30. The correct answer is (b). Some of the other responses can *contribute* to bias, but do not represent what is *meant* by bias.
31. The correct answer is (d).  $\mu_x = 1(0.10) + 2(0.20) + 3(0.40) + 4(0.10) + 5(0.20) = 3.1$ .
32. The correct answer is (e).  $P(\text{at least one 4}) = \binom{8}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^7 + \binom{8}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^6 + \dots + \binom{8}{8} \left(\frac{1}{6}\right)^8 \left(\frac{5}{6}\right)^0 = 1 - P(0 \text{ fours}) = 1 - \binom{8}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^8$
33. The correct answer is (b). When  $H_0$  is false, power =  $1 - \beta$ , where  $\beta$  is the probability of making a type II error (the mistake of failing to reject a false null hypothesis).  $1 - \alpha$  is the probability of correctly failing to reject a true null hypothesis. (a) and (c) make it easier to reject the null hypothesis, thus increasing power. (c) is essentially a definition of power. (e) is true since you can increase power by increasing the difference between the real and hypothesized value of the parameter.
34. The correct answer is (d). You must have 4 out of the 10 possibilities. (d) is the only one that accomplishes this.
35. The correct answer is (d). The opposite of III is correct: the greater the number of degrees of freedom, the *less* the variance of a  $t$ -distribution.

36. The correct answer is (d). It isn't true because this is an observational study and simply shows an association, not a causal relationship.
37. The correct answer is (e) since we are told that there is evidence that temperature affects distance traveled. (a) is incorrect since the type of golf ball is the treatment variable, not a blocking variable. (b) is incorrect since there is no reason to block by size since we have no evidence that size is related to distance. (c) is incorrect for the same reason as (b). (d) is incorrect since there is reason to believe that some variable not under study is related to the outcomes to be measured.
38. The correct answer is (a). From the cumulative frequencies, we can determine the relative frequency of each value.

VALUE (X)	CUMULATIVE FREQUENCY	RELATIVE FREQUENCY
2	0.15	0.15
3	0.25	0.10
5	0.45	0.20
7	0.95	0.50
10	1.00	0.05

Then,  $\mu_x = 2(0.15) + 3(0.10) + 5(0.20) + 7(0.50) + 10(0.05) = 5.6$ . The median is the value at the 50th percentile. Because only 45% of the scores are at 5 or below, and 95% of the scores are at 7 or below, the median must be at 7.

39. The correct answer is (a). Since there are 1000 scores, the five-number summary will occur at the first (minimum), 250th (25th percentile), 500th (median), 750th (75th percentile), and 1000th (maximum) scores. Drawing a horizontal line from each of these values on the vertical axis to the cumulative frequency graph and then dropping down to the horizontal axis yields, most closely, the values in (a).
40. The correct answer is (e). We are 95% confident (most polls use a 95% confidence interval) that the true population proportion of voters who intend to vote for Sleazy is in the interval  $46\% \pm 3\%$  or between 43% and 49%. This is equivalent to saying the proportion is *likely* to be between 43% and 49%.

### Solutions to Practice Exam 2, Section II, Part A

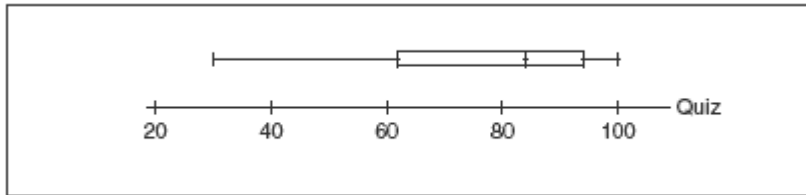
#### Solution to #1

1. The median is the middle score. Because there are 20 scores, the median lies between the 10th and 11th scores as shown on the stemplot.

3		0
4		
5		226
6		2238
7		5
8		4789
9		135578
10		0

$$\frac{84 + 87}{2} = 85.5$$

The 10th and 11th scores are 84 and 87. Thus the median is



- 2.
3.  $1.5(\text{IQR}) = 1.5(94 - 62) = 1.5(32) = 48$ . Outliers would be values beyond  $62 - 48 = 14$  or  $94 + 48 = 142$ . The lowest score a student could get that would not be considered an outlier would be 14.

**Solution to #2**

1. Number =  $229.402 - 7.057(\text{birth month})$
2. Each month, from January to December, the average draft number is predicted to decrease by 7.057.
3. Yes. The  $t$ -statistic of  $-5.49$  is the test statistic for the hypothesis  $H_0: \beta = 0$  against the alternative  $H_A: \beta \neq 0$ . This  $t$ -statistic tells us that the  $P$ -value is approximately 0. This is very strong evidence against the null hypothesis, which supports the contention that there is a predictive relationship between birth month and average draft number. The conclusion is that the drawing is biased against those born later in the year. That is, those born later in the year are more likely to get drafted than those born early in the year.
4.  $df = n - 2 = 12 - 2 = 10 \Rightarrow t^* = 2.228$  (from Table B; if you have a TI-84 with the invT function,  $\text{invT}(0.975, 10) = 2.228$ ). The 95% confidence interval is  $-7.057 \pm 2.228(1.286) = (-9.92, -4.19)$ . We are 95% confident that the true slope of the regression line is between  $-9.92$  and  $-4.19$ . Note that this is consistent with our finding in part (c) that the regression line has a nonzero slope.

**Solution to #3**

1. It is an observational study. The researcher is not controlling the treatments to the two different groups in the study. Rather, the groups are self-selected based on where they bought their mattresses. The researcher has simply observed and recorded outcomes. In an experiment, the researcher would have had control over which volunteers slept on which mattress.
2. A confounding variable would be a variable that differentially affects one group of buyers more than the other. For example, a possible confounding variable might be that people who buy their mattresses at an upscale location are better off financially than those who buy for economy. They might be more content and less stressed as a group, and this could affect the quality of their sleep.
3. No. You cannot infer a cause-and-effect relationship from an observational study because of the possible presence of confounding variables as described in (b). All you know is that the groups experienced different qualities of sleep. You do not know why.

**Solution to #4**

We note that the scores are paired by school, so we will want to use a matched pairs test.

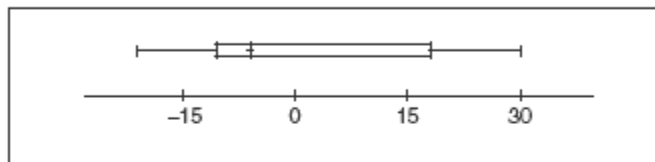
Let  $\mu_d$  be the true mean difference between the graduation rates of all female students and female basketball players.

$$H_0: \mu_d = 0.$$

$$H_A: \mu_d \neq 0.$$

We need to use the paired differences for analysis. Adding a column to the table in the problem with the differences, we have:

SCHOOL	ALL	BASKETBALL PLAYERS	DIFFERENCES ( $d$ )
A	59	67	-8
B	54	75	-21
C	39	52	-13
D	71	55	16
E	69	40	29
F	74	55	19
G	64	67	-3
H	70	40	30
I	40	50	-10
J	85	95	-10



The problem states that the sample is a random sample. A boxplot of the data shows some skewness but no outliers. We can proceed with a one-sample  $t$ -test.

$$n = 10 \Rightarrow df = 10 - 1 = 9. \quad \bar{x}_d = 2.9, \quad s_d = 18.73$$

$$t = \frac{2.9 - 0}{18.73 / \sqrt{10}} = 0.49 \Rightarrow P\text{-value} > 0.50 \text{ (from Table B; using the TI-83/84,}$$

$$P\text{-value} = 2 \times \text{tcdf}(0.49, 9) = 0.636).$$

Because the  $P$ -value is quite large, we have no evidence to reject the null and conclude that there is a significant difference between the graduation rates of all female college students and female basketball players. (A good research follow-up at this point would be to do the same analysis on *male* basketball players to see if a difference in graduation rates exists there.)

#### Solution to #5

1.  $Time = 163.1 - 1.23(25) = 132.35$ . The predicted time for 1999 is about 2 hours 12 minutes.
2. Residual = actual - predicted =  $143 - 132 = 11$  minutes.
3. Not really. The residual plot shows a pattern for underestimating the actual times at the beginning and end of the 20-year period and overestimating the actual times during the middle years.
4. The scatterplot for the raw data appears to be more linear beginning about year 5. Because the goal is to predict beyond the data, a line might be a better model beginning with year 5 rather than with year 1. That is, redo the regression beginning in 1980 rather than in 1975.

## Solutions to Practice Exam 2, Section II, Part B

### Solution to #6

# Faces	0	1	2	3
Win ( $x$ )	-1	1	2	3
$P(x)$	$\binom{3}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3$ = 0.579	$\binom{3}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^2$ = 0.347	$\binom{3}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^1$ = 0.069	$\binom{3}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0$ = 0.005

- 
- $\mu_X = -1(0.579) + 1(0.347) + 2(0.069) + 3(0.005) = -0.079$ . Your expected winnings on each \$1 bet is  $-7.9\text{¢}$ .  
(Remember that this can be done on the TI-83/84 as follows: put the values of  $X$  ( $\{-1, 1, 2, 3\}$ ) in L1 and the values of  $P(X)$  ( $\{0.579, 0.347, 0.069, 0.005\}$ ) in L2. Then do STAT CALC 1-VAR STATS L1,L2 to find  $\mu_x$ , which is given as  $\bar{X}$  on the calculator.)

$$\sigma_X = \sqrt{(-1 - (-0.079))^2(0.579) + (1 + 0.079)^2(0.347) + (2 + 0.079)^2(0.069) + (3 + 0.079)^2(0.005)} = 1.114.$$

(This value can be found on the TI-83/84 as indicated above by putting the  $X$  values in L1, the  $P(X)$  in L2, and doing STAT CALC 1-VAR STATS L1,L2.)

- With 100 trials, you can assume that the central limit theorem kicks in and that the sampling distribution of  $\bar{X}$  is approximately normally distributed. Let  $\bar{x}$  be your average winning on 100 plays of the game. Then

$$\sigma_{\bar{x}} = \frac{1.114}{\sqrt{100}} = 0.111 \quad P(z > \frac{0 - (-0.079)}{0.111} = 0.712) = 1 - 0.7611 = 0.2389 \text{ (from Table A; n the TI-83/84: normalcdf}(0.712, 100) = 0.239).$$

You have about a 24% chance of making a profit (or a 76% chance of losing money) after 100 plays.

$$P(\bar{x} > 0) = P(z > \frac{0 - (-0.079)}{1.11/\sqrt{10000}} = 7.12) = 0$$

For 10,000 plays, (normalcdf (7.12, 100) = 5.434E-13, or  $5.434 \times 10^{-13}$ .) You have essentially no chance of making a profit after 10,000 plays. The casino is virtually guaranteed to make money and the players, on average, are guaranteed to lose.

## Scoring Sheet for Practice Exam 2

### Section I: Multiple Choice

$$\left[ \frac{\text{number correct}}{\text{(out of 40)}} - \left( \frac{1}{4} \times \frac{\text{number wrong}}{\text{number wrong}} \right) \right] \times 1.25 = \frac{\text{multiple-choice score (if less than zero, enter zero)}}{\text{weighted section I score (do not round)}}$$

### Section II: Free Response

$$\text{Question 1 } \frac{\text{_____}}{\text{(out of 4)}} \times 1.875 = \frac{\text{_____}}{\text{(do not round)}}$$

$$\text{Question 2 } \frac{\text{_____}}{\text{(out of 4)}} \times 1.875 = \frac{\text{_____}}{\text{(do not round)}}$$

$$\text{Question 3 } \frac{\text{_____}}{\text{(out of 4)}} \times 1.875 = \frac{\text{_____}}{\text{(do not round)}}$$

$$\text{Question 4 } \frac{\text{_____}}{\text{(out of 4)}} \times 1.875 = \frac{\text{_____}}{\text{(do not round)}}$$

$$\text{Question 5 } \frac{\text{_____}}{\text{(out of 4)}} \times 1.875 = \frac{\text{_____}}{\text{(do not round)}}$$

$$\text{Question 6 } \frac{\text{_____}}{\text{(out of 4)}} \times 3.125 = \frac{\text{_____}}{\text{(do not round)}}$$

$$\text{Sum} = \frac{\text{_____}}{\text{weighted section II score (do not round)}}$$

### Composite Score

$$\frac{\text{_____}}{\text{weighted section I score}} + \frac{\text{_____}}{\text{weighted section II score}} = \frac{\text{_____}}{\text{composite score (round to nearest whole number)}}$$