Practice Midterm #4

Section I: Multiple Choice *Choose the best answer*

1. You look at real estate ads for houses in Sarasota, Florida. Many houses range from \$200,000 to \$400,000 in price. The few houses on the water, however, have prices up to \$15 million. Which of the following statements best describes the distribution of home prices in Sarasota?

(a) The distribution is most likely skewed to the left, and the mean is greater than the median.

(b) The distribution is most likely skewed to the left, and the mean is less than the median.

(c) The distribution is roughly symmetric with a few high outliers, and the mean is approximately equal to the median.

(d) The distribution is most likely skewed to the right, and the mean is greater than the median.

(e) The distribution is most likely skewed to the right, and the mean is less than the median.

2. A child is 40 inches tall, which places her at the 90th percentile of all children of similar age. The heights for children of this age form an approximately Normal distribution with a mean of 38 inches. Based on this information, what is the standard deviation of the heights of all children of this age?

(a) 0.20 inches (b) 0.31 inches (c) 0.65 inches (d) 1.21 inches (e) 1.56 inches

3. A large set of test scores has mean 60 and standard deviation 18. If each score is doubled, and then 5 is subtracted from the result, the mean and standard deviation of the new scores are

(a) mean 115; std. dev. 31. (b) mean 115; std. dev. 36. (c) mean 120; std. dev. 6.

(d) mean 120; std. dev. 31. (e) mean 120; std. dev. 36.

4. For a certain experiment, the available experimental units are eight rats, of which four are female (F1, F2, F3, F4) and four are male (M1, M2, M3, M4). There are to be four treatment groups, A, B, C, and D. If a randomized block design is used, with the experimental units blocked by gender, which of the following assignments of treatments is impossible?

(a) A → (F1, M1), B → (F2, M2), C → (F3, M3), D → (F4, M4)

(b) A → (F1, M2), B → (F2, M3), C → (F3, M4), D → (F4, M1)

(c) A → (F1, M2), B → (F3, F2), C → (F4, M1), D → (M3, M4)

(d) A → (F4, M1), B → (F2, M3), C → (F3, M2), D → (F1, M4)

(e) A → (F4, M1), B → (F1, M4), C → (F3, M2), D → (F2, M3)

5. For a biology project, you measure the weight in grams (g) and the tail length in millimeters (mm) of a group of mice. The equation of the least-squares line for predicting tail length from weight is

predicted tail length = 20 + 3 × weight Which of the following is *not* correct?

(a) The slope is 3, which indicates that a mouse's weight should increase by about 3 grams for each additional millimeter of tail length.

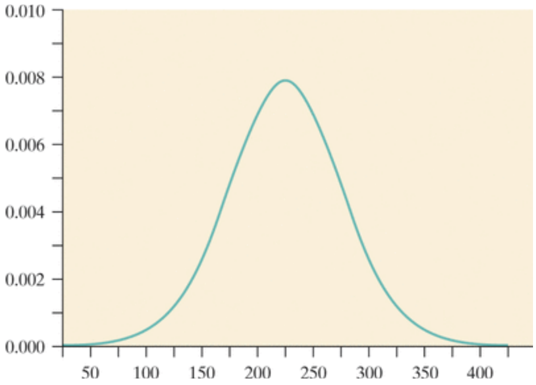
(b) The predicted tail length of a mouse that weighs 38 grams is 134 millimeters.

(c) By looking at the equation of the least-squares line, you can see that the correlation between weight and tail length is positive.

(d) If you had measured the tail length in centimeters instead of millimeters, the slope of the regression line would have been 3/10 = 0.3.

(e) One mouse weighed 29 grams and had a tail length of 100 millimeters. The residual for this mouse is −7.

6. The figure below shows a Normal density curve. Which of the following gives the best estimates for the mean and standard deviation of this Normal distribution?



(a) $\mu = 200, \sigma = 50$ (b) $\mu = 200, \sigma = 25$ (c) $\mu = 225, \sigma = 50$ (d) $\mu = 225, \sigma = 25$ (e) $\mu = 225, \sigma = 275$

7. The owner of a chain of supermarkets notices that there is a positive correlation between the sales of beer and the sales of ice cream over the course of the previous year. During seasons when sales of beer were above average, sales of ice cream also tended to be above average. Likewise, during seasons when sales of beer were below average, sales of ice cream also tended to be below average. Which of the following would be a valid conclusion from these facts?

(a) Sales records must be in error. There should be no association between beer and ice cream sales.

(b) Evidently, for a significant proportion of customers of these supermarkets, drinking beer causes a desire for ice cream or eating ice cream causes a thirst for beer.

(c) A scatterplot of monthly ice cream sales versus monthly beer sales would show that a straight line describes the pattern in the plot, but it would have to be a horizontal line.

(d) There is a clear negative association between beer sales and ice cream sales.

(e) The positive correlation is most likely a result of the lurking variable temperature; that is, as temperatures increase, so do both beer sales and ice cream sales.

8. Here are the IQ scores of 10 randomly chosen fifth-grade students:

145 139 126 122 125 130 96 110 118 118 Which of the following statements about this data set is *not* true?

(a) The student with an IQ of 96 is considered an outlier by the 1.5 × *IQR* rule.

(b) The five-number summary of the 10 IQ scores is 96, 118, 123.5, 130, 145.

(c) If the value 96 were removed from the data set, the mean of the remaining 9 IQ scores would be higher than the mean of all 10 IQ scores.

(d) If the value 96 were removed from the data set, the standard deviation of the remaining 9 IQ scores would be lower than the standard deviation of all 10 IQ scores.

(e) If the value 96 were removed from the data set, the *IQR* of the remaining 9 IQ scores would be lower than the *IQR* of all 10 IQ scores.

9. Before he goes to bed each night, Mr. Kleen pours dishwasher powder into his dishwasher and turns it on. Each morning, Mrs. Kleen weighs the box of dishwasher powder. From an examination of the data, she concludes that Mr. Kleen dispenses a rather consistent amount of powder each night. Which of the following statements is true?

I. There is a high positive correlation between the number of days that have passed since the box of dishwasher powder was opened and the amount of powder left in the box.

II. A scatterplot with days since purchase as the explanatory variable and amount of dishwasher powder used as the response variable would display a strong positive association.

III. The correlation between the amount of powder left in the box and the amount of powder used should be −1.

(a) I only (b) II only (c) III only (d) II and III only (e) I, II, and III

10. The General Social Survey (GSS), conducted by the National Opinion Research Center at the University of Chicago, is a major source of data on social attitudes in the United States. Once each year, 1500 adults are interviewed in their homes all across the country. The subjects are asked their opinions about sex and marriage, attitudes toward women, welfare, foreign policy, and many other issues. The GSS begins by selecting a sample of counties from the 3000 counties in the country. The counties are divided into urban, rural, and suburban; a separate sample is chosen at random from each group. This is a

(a) simple random sample. (b) systematic random sample. (c) cluster sample.

(d) stratified random sample. (e) voluntary response sample.

11. You are planning an experiment to determine the effect of the brand of gasoline and the weight of a car on gas mileage measured in miles per gallon. You will use a single test car, adding weights so that its total weight is 3000, 3500, or 4000 pounds. The car will drive on a test track at each weight using each of Amoco, Marathon, and Speedway gasoline. Which is the best way to organize the study?

(a) Start with 3000 pounds and Amoco and run the car on the test track. Then do 3500 and 4000 pounds. Change to Marathon and go through the three weights in order. Then change to Speedway and do the three weights in order once more.

(b) Start with 3000 pounds and Amoco and run the car on the test track. Then change to Marathon and then to Speedway without changing the weight. Then add weights to get 3500 pounds and go through the three gasolines in the same order. Then change to 4000 pounds and do the three gasolines in order again.

(c) Choose a gasoline at random, and run the car with this gasoline at 3000, 3500, and 4000 pounds in order. Choose one of the two remaining gasolines at random and again run the car at 3000, then 3500, then 4000 pounds. Do the same with the last gasoline.

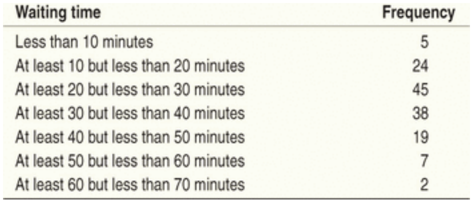
(d) There are nine combinations of weight and gasoline. Run the car several times using each of these combinations. Make all these runs in random order.

(e) Randomly select an amount of weight and a brand of gasoline, and run the car on the test track. Repeat this process a total of 30 times.

12. A linear regression was performed using the five following data points: A(2, 22), B(10, 4), C(6, 14), D(14, 2), E(18, −4). The residual for which of the five points has the largest absolute value?

(a) A (b) B (c) C (d) D (e) E

13. The frequency table below summarizes the times in the last month that patients at the emergency room of a small-city hospital waited to receive medical attention.



Waiting time	Frequency
Less than 10 minutes	5
At least 10 but less than 20 minutes	24
At least 20 but less than 30 minutes	45
At least 30 but less than 40 minutes	38
At least 40 but less than 50 minutes	19
At least 50 but less than 60 minutes	7
At least 60 but less than 70 minutes	2

Which of the following represents possible values for the median and mean waiting times for the emergency room last month?

(a) median = 27 minutes and mean = 24 minutes

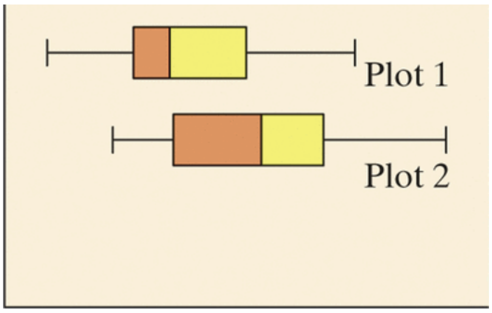
(b) median = 28 minutes and mean = 30 minutes

(c) median = 31 minutes and mean = 35 minutes

(d) median = 35 minutes and mean = 39 minutes

(e) median = 45 minutes and mean = 46 minutes

14. Boxplots of two data sets are shown.



Based on the boxplots, which statement below is true?

(a) The spread of both plots is about the same.

(b) The means of both plots are approximately equal.

(c) Plot 2 contains more data points than Plot 1.

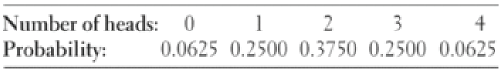
(d) The medians are approximately equal.

(e) Plot 1 is more symmetric than Plot 2.

15. The five-number summary for a data set is given by min = 5, Q_1 = 18, M = 20, Q_3 = 40, max = 75. If you wanted to construct a modified boxplot for the data set (that is, one that would show outliers, if any existed), what would be the maximum possible length of the right-side "whisker"?

(a) 33 (b) 35 (c) 45 (d) 53 (e) 55

16. The probability distribution for the number of heads in four tosses of a coin is given by



Number of heads:	0	1	2	3	4
Probability:	0.0625	0.2500	0.3750	0.2500	0.0625

The probability of getting at least one *tail* in four tosses of a coin is

(a)0.2500. (b) 0.3125. (c) 0.6875. (d) 0.9375. (e) none of these.

17. In a certain large population of adults, the distribution of IQ scores is strongly left-skewed with a mean of 122 and a standard deviation of 5. Suppose 200 adults are randomly selected from this population for a market research study. The distribution of the sample mean of IQ scores is

(a) left-skewed with mean 122 and standard deviation 0.35.

(b) exactly Normal with mean 122 and standard deviation 5.

(c) exactly Normal with mean 122 and standard deviation 0.35.

(d) approximately Normal with mean 122 and standard deviation 5.

(e) approximately Normal with mean 122 and standard deviation 0.35.

18. A 10-question multiple-choice exam offers 5 choices for each question. Jason just guesses the answers, so he has probability 1/5 of getting any one answer correct. You want to perform a simulation to determine the number of correct answers that Jason gets. One correct way to use a table of random digits to do this is the following:

(a) One digit from the random digit table simulates one answer, with 5 = right and all other digits = wrong. Ten digits from the table simulate 10 answers.

(b) One digit from the random digit table simulates one answer, with 0 or 1 = right and all other digits = wrong. Ten digits from the table simulate 10 answers.

(c) One digit from the random digit table simulates one answer, with odd = right and even = wrong. Ten digits from the table simulate 10 answers.

(d) Two digits from the random digit table simulate one answer, with 00 to 20 = right and 21 to 99 = wrong. Ten pairs of digits from the table simulate 10 answers.

(e) Two digits from the random digit table simulate one answer, with 00 to 05 = right and 06 to 99 = wrong. Ten pairs of digits from the table simulate 10 answers.

19. Suppose we roll a fair die four times. The probability that a 6 occurs on exactly one of the rolls is

(a) $4\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$ (b) $\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$ (c) $4\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$ (d) $\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$ (e) $6\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$

20. You want to take an SRS of 50 of the 816 students who live in a dormitory on a college campus. You label the students 001 to 816 in alphabetical order. In the table of random digits, you read the entries

95592 94007 69769 33547 72450 16632 81194 14873

The first three students in your sample have labels

(a) 955, 929, 400. (b) 400, 769, 769. (c) 559, 294, 007. (d) 929, 400, 769. (e) 400, 769, 335.

21. The number of unbroken charcoal briquets in a twenty-pound bag filled at the factory follows a Normal distribution with a mean of 450 briquets and a standard deviation of 20 briquets. The company expects that a certain number of the bags will be underfilled, so the company will replace for free the 5% of bags that have too few briquets. What is the minimum number of unbroken briquets the bag would have to contain for the company to avoid having to replace the bag for free?

(a) 404 (b) 411 (c) 418 (d) 425 (e) 448

22. You work for an advertising agency that is preparing a new television commercial to appeal to women. You have been asked to design an experiment to compare the effectiveness of three versions of the commercial. Each subject will be shown one of the three versions and then asked about her attitude toward the product. You think there may be large differences between women who are employed and those who are not. Because of these differences, you should use

(a) a block design, but not a matched pairs design. (b) a completely randomized design.

(c) a matched pairs design. (d) a simple random sample. (e) a stratified random sample.

23. Suppose that you have torn a tendon and are facing surgery to repair it. The orthopedic surgeon explains the risks to you. Infection occurs in 3% of such operations, the repair fails in 14%, and both infection and failure occur together 1% of the time. What is the probability that the operation is successful for someone who has an operation that is free from infection?

(a) 0.0767 (b) 0.8342 (c) 0.8400 (d) 0.8660 (e) 0.9900

24. Social scientists are interested in the association between high school graduation rate (HSGR) and the percent of U.S. families living in poverty (POV). Data were collected from all 50 states and the District of Columbia, and a regression analysis was conducted. The resulting least-squares regression line is given by $\widehat{\text{POV}} = 59.2 - 0.620(\text{HSGR})$ with $r^2 = 0.802$. Based on the information, which of the following is the best interpretation for the slope of the least-squares regression line?

(a) For each 1% increase in the graduation rate, the percent of families living in poverty is predicted to decrease by approximately 0.896.

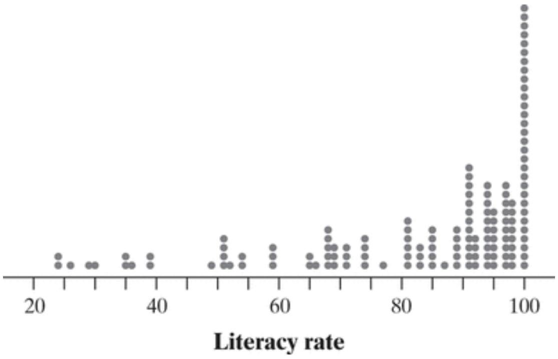
(b) For each 1% increase in the graduation rate, the percent of families living in poverty is predicted to decrease by approximately 0.802.

(c) For each 1% increase in the graduation rate, the percent of families living in poverty is predicted to decrease by approximately 0.620.

(d) For each 1% increase in the percent of families living in poverty, the graduation rate is predicted to increase by approximately 0.802.

(e) For each 1% increase in the percent of families living in poverty, the graduation rate is predicted to decrease by approximately 0.620.

Here is a dotplot of the adult literacy rates in 177 countries in 2008, according to the United Nations. For example, the lowest literacy rate was 23.6%, in the African country of Burkina Faso. Use the dotplot below to answer Questions 25 to 27.



25. The overall shape of this distribution is

(a) clearly skewed to the right. (b) clearly skewed to the left. (c) roughly symmetric.

(d) uniform. (e) There is no clear shape.

26. The mean of this distribution (*don't* try to find it) will be

(a) very close to the mode. (b) greater than the median. (c) less than the median.

(d) You can't say, because the median is random. (e) You can't say, because the mean is random.

27. Based on the shape of this distribution, what numerical measures would best describe it?

(a) The five-number summary (b) The mean and standard deviation (c) The mean and the quartiles

(d) The median and the standard deviation (e) It is not possible to determine which numerical values to use.

28. The correlation between the age and height of children under the age of 12 is found to be $r = 0.60$. Suppose we use the age x of a child to predict the height y of the child. What can we conclude?

(a)The height is generally 60% of a child's weight.

(b) About 60% of the time, age will accurately predict height.

(c) The fraction of the variation in heights explained by the least-squares regression line of y on x is 0.36.

(d) The least-squares regression line of y on x has a slope of 0.6.

(e) Thirty-six percent of the time, the least-squares regression line accurately predicts height.

29. An agronomist wants to test three different types of fertilizer (A, B, and C) on the yield of a new variety of wheat. The yield will be measured in bushels per acre. Six one-acre plots of land were randomly assigned to each of the three fertilizers. The treatment, experimental unit, and response variable are, respectively,

(a) a specific fertilizer, bushels per acre, a plot of land. (b) a plot of land, bushels per acre, a specific fertilizer.

(c) random assignment, a plot of land, bushels per acre. (d) a specific fertilizer, a plot of land, bushels per acre.

(e) a specific fertilizer, the agronomist, bushels per acre.

30. Which one of the following would be a correct interpretation if you have a *z*-score of +2.0 on an exam?

(a) It means that you missed two questions on the exam.

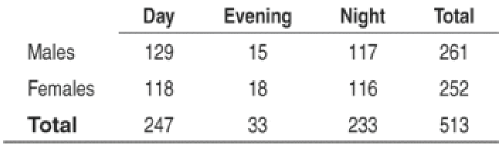
(b) It means that you got twice as many questions correct as the average student.

(c) It means that your grade was two points higher than the mean grade on this exam.

(d) It means that your grade was in the upper 2% of all grades on this exam.

(e) It means that your grade is two standard deviations above the mean for this exam.

31. Records from a random sample of dairy farms yielded the information below on the number of male and female calves born at various times of the day.



	Day	Evening	Night	Total
Males	129	15	117	261
Females	118	18	116	252
Total	247	33	233	513

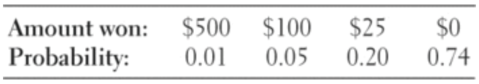
What is the probability that a randomly selected calf was born in the night or was a female?

(a) $\frac{369}{513}$ (b) $\frac{485}{513}$ (c) $\frac{116}{513}$ (d) $\frac{116}{252}$ (e) $\frac{116}{233}$

32. When people order books from a popular online source, they are shipped in standard-sized boxes. Suppose that the mean weight of the boxes is 1.5 pounds with a standard deviation of 0.3 pounds, the mean weight of the packing material is 0.5 pounds with a standard deviation of 0.1 pounds, and the mean weight of the books shipped is 12 pounds with a standard deviation of 3 pounds. Assuming that the weights are independent, what is the standard deviation of the total weight of the boxes that are shipped from this source?

(a) 1.84 (b) 2.60 (c) 3.02 (d) 3.40 (e) 9.10

33. A grocery chain runs a prize game by giving each customer a ticket that may win a prize when the box is scratched off. Printed on the ticket is a dollar value ($500, $100, $10) or the statement, “This ticket is not a winner.” Monetary prizes can be redeemed for groceries at the store. Here are the distribution of the prize values and the associated probabilities for each prize:

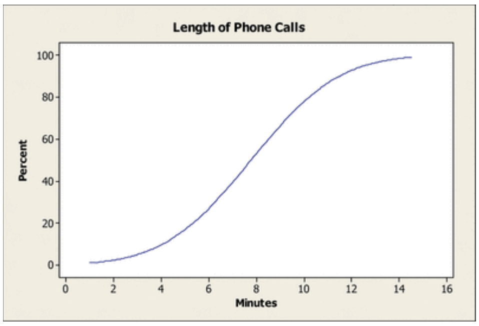


Amount won:	$500	$100	$25	$0
Probability:	0.01	0.05	0.20	0.74

Which of the following are the mean and standard deviation, respectively, of the winnings?

(a) $15.00, $2900.00 (b) $15.00, $53.85 (c) $15.00,$53.85 (d) $156.25, $53.85 (e) $156.25, $26.93

34. A large company is interested in improving the efficiency of its customer service and decides to examine the length of the business phone calls made to clients by its sales staff. A cumulative relative frequency graph is shown below from data collected over the past year. According to the graph, the shortest 80% of calls will take how long to complete?



(a) Less than 10 minutes. (b) At least 10 minutes. (c) Exactly 10 minutes.

(d) At least 5.5 minutes. (e) Less than 5.5 minutes.

Section II: Free Response *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 and completeness of your results and explanations.*

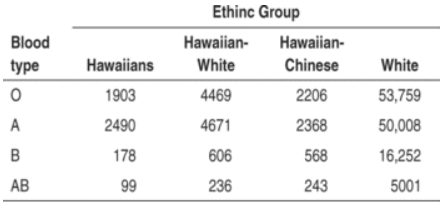
1. A health worker is interested in determining if omega-3 fish oil can help reduce cholesterol in adults. She obtains permission to examine the health records of 200 people in a large medical clinic and classifies them according to whether or not they take omega-3 fish oil. She also obtains their latest cholesterol readings and finds that the mean cholesterol reading for those who are taking omega-3 fish oil is 18 points lower than the mean for the group not taking omega-3 fish oil.

(a) Is this an observational study or an experiment? Explain.

(b) Do these results provide convincing evidence that taking omega-3 fish oil lowers cholesterol?

(c) Explain the concept of confounding in the context of this study and give one example of a possible confounding variable.

2. There are four major blood types in humans: O, A, B, and AB. In a study conducted using blood specimens from the Blood Bank of Hawaii, individuals were classified according to blood type and ethnic group. The ethnic groups were Hawaiian, Hawaiian-White, Hawaiian-Chinese, and White. Suppose that a blood bank specimen is selected at random.



	Ethinc Group			
Blood type	Hawaiians	Hawaiian-White	Hawaiian-Chinese	White
O	1903	4469	2206	53,759
A	2490	4671	2368	50,008
B	178	606	568	16,252
AB	99	236	243	5001

(a) Find the probability that the specimen contains type O blood or comes from the Hawaiian-Chinese ethnic group. Show your work.

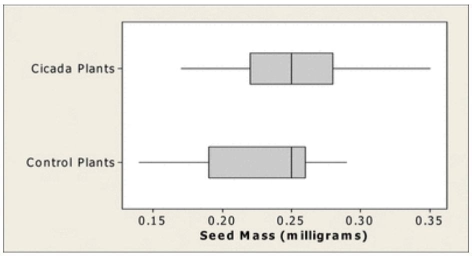
(b) What is the probability that the specimen contains type AB blood, given that it comes from the Hawaiian ethnic group? Show your work.

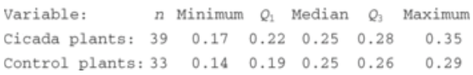
(c) Are the events “type B blood” and “Hawaiian ethnic group” independent? Give appropriate statistical evidence to support your answer.

Now suppose that two blood bank specimens are selected at random.

(d)Find the probability that at least one of the specimens contains type A blood from the White ethnic group.

3. Every 17 years, swarms of cicadas emerge from the ground in the eastern United States, live for about six weeks, and then die. (There are several different “broods,” so we experience cicada eruptions more often than every 17 years.) There are so many cicadas that their dead bodies can serve as fertilizer and increase plant growth. In a study, a researcher added 10 cicadas under 39 randomly selected plants in a natural plot of American bellflowers on the forest floor, leaving other plants undisturbed. One of the response variables measured was the size of seeds produced by the plants. Here are the boxplots and summary statistics of seed mass (in milligrams) for 39 cicada plants and 33 undisturbed (control) plants:





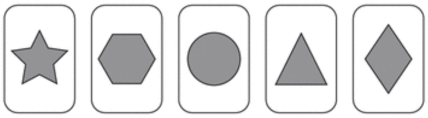
```
Variable:         n Minimum  Q1  Median  Q3   Maximum
Cicada plants:  39   0.17   0.22  0.25  0.28    0.35
Control plants: 33   0.14   0.19  0.25  0.26    0.29
```

(a) Is this an observational study or an experiment? Explain.

(b) Based on the graphical displays, which distribution has the larger mean? Justify your answer.

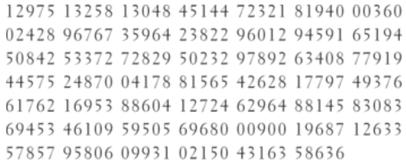
(c) Do the data support the idea that dead cicadas can serve as fertilizer? Give graphical and numerical evidence to support your conclusion.

4. Five cards, each with a different symbol, are shuffled and you choose one. If it is the diamond, you win $5.00. The cards are reshuffled after each draw. You must pay $1.00 for each selection. You continue to play until you select the diamond. Is this a fair game (that is, on average will you win the same amount as you lose)?



(a) Describe how you will carry out a simulation of this game using the random digit table below. Be sure to indicate what the digits will represent.

(b) Perform 10 repetitions of your simulation. Copy the random digit table onto your paper. Mark on or above the table so that someone can follow your work.

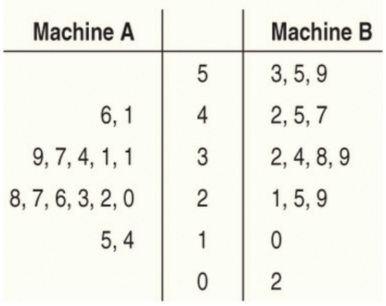


```
12975 13258 13048 45144 72321 81940 00360
02428 96767 35964 23822 96012 94591 65194
50842 53372 72829 50232 97892 63408 77919
44575 24870 04178 81565 42628 17797 49376
61762 16953 88604 12724 62964 88145 83083
69453 46109 59505 69680 00900 19687 12633
57857 95806 09931 02150 43163 58636
```

(c) Based on your simulation, what is the average number of cards you would need to draw in order to obtain a diamond? Justify your answer.

(d) Is this a fair game (that is, on average will you win the same amount as you lose)? Explain your reasoning.

5. The manufacturer of exercise machines for fitness centers has designed two new elliptical machines that are meant to increase cardiovascular fitness. The two machines are being tested on 30 volunteers at a fitness center near the company's headquarters. The volunteers are randomly assigned to one of the machines and use it daily for two months. A measure of cardiovascular fitness is administered at the start of the experiment and again at the end. The following table contains the differences in the two scores (After − Before) for the two machines. Note that higher scores indicate larger gains in fitness.



Machine A		Machine B
	5	3, 5, 9
6, 1	4	2, 5, 7
9, 7, 4, 1, 1	3	2, 4, 8, 9
8, 7, 6, 3, 2, 0	2	1, 5, 9
5, 4	1	0
	0	2

(a) Write a few sentences comparing the distributions of cardiovascular fitness gains from the two elliptical machines.

(b) Which machine should be chosen if the company wants to advertise it as achieving the highest overall gain in cardiovascular fitness? Explain your reasoning.

(c) Which machine should be chosen if the company wants to advertise it as achieving the most consistent gain in cardiovascular fitness? Explain your reasoning.

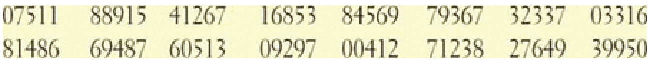
(d) Give one reason why the advertising claims of the company (the scope of inference) for this experiment would be limited. Explain how the company could broaden that scope of inference.

6. Those who advocate for monetary incentives in a work environment claim that this type of incentive has the greatest appeal because it allows the winners to do what they want with their winnings. Those in favor of tangible incentives argue that money lacks the emotional appeal of, say, a weekend for two at a romantic country inn or elegant hotel or a weeklong trip to Europe.

A few years ago a national tire company, in an effort to improve sales of a new line of tires, decided to test which method—offering cash incentives or offering non-cash prizes such as vacations—was more successful in increasing sales. The company had 60 retail sales districts of various sizes across the country and data on the previous sales volume for each district.

(a) Describe a completely randomized design using the 60 retail sales districts that would help answer this question.

(b) Explain how you would use the table of random digits below to do the randomization that your design requires. Then use your method to assign treatments to the first 3 experimental units. Show your work clearly.

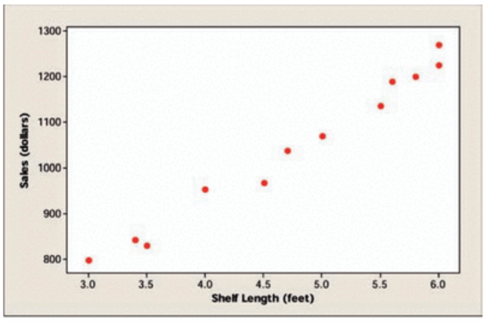


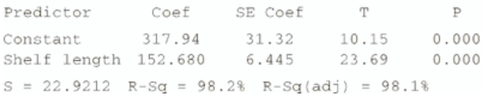
07511	88915	41267	16853	84569	79367	32337	03316
81486	69487	60513	09297	00412	71238	27649	39950

(c) One of the company's officers suggested that it would be better to use a matched pairs design instead of a completely randomized design. Explain how you would change your design to accomplish this.

7. In retail stores, there is a lot of competition for shelf space. Not only are there national brands for most products, but many stores have their own in-house brands. Since shelf space is not infinite, the question is how many linear feet to allocate to each product and which shelf (top, bottom, or somewhere in the middle) to put it on. The middle shelf is the most popular and lucrative, since many shoppers, if undecided, will simply pick the product that is at eye level.

A local store that sells many upscale goods is trying to determine how much shelf space to allocate to its own brand of men's personal-grooming products. The middle shelf space is randomly varied between three and six linear feet over the next 12 weeks, and weekly sales revenue (in dollars) from the store's brand of personal-grooming products for men is recorded. Below is some computer output from the study, along with a scatterplot.





```
Predictor       Coef     SE Coef      T        P

Constant       317.94     31.32     10.15    0.000
Shelf length  152.680     6.445     23.69    0.000

S = 22.9212  R-Sq = 98.2%  R-Sq(adj) = 98.1%
```

(a) Does it appear that the weekly sales revenue is related to the shelf length allocated to the house brand? Justify your answer.

(b) Write the equation of the least-squares regression line. Be sure to define any variables you use.

(c) If the store manager were to decide to allocate five linear feet of shelf space to the store's brand of men's grooming products, what is the best estimate of the weekly sales revenue?

(d) Interpret the value of *s*.

(e) Identify and interpret the coefficient of determination.

(f) The store manager questions the intercept of the regression line: "Am I supposed to believe that this analysis tells me that I can sell these products with no shelf space?" How do you answer her?

Answers

1. D

2. E

3. B

4. C

5. A

6. C

7. E

8. E

9. D

10. D

11. D

12. B

13. B

14. A

15. A

16. D

17. E

18. B

19. C

20. E

21. C

22. A

23. D

24. C

25. B

26. C

27. A

28. C

29. D

30. E

31. A

32. C

33. B

34. A

1. **(a)** This is an observational study. No treatments were imposed on the subjects. **(b)** No. Since this was an observational study and not a randomized controlled experiment, no cause-and-effect conclusions are possible. **(c)** Two variables are confounded when their effects on the cholesterol level cannot be distinguished from one another. For example, people who take omega-3 fish oil might also be more health conscious in general and do other things such as eat more healthfully or exercise more. Researchers would not know whether it was the omega-3 fish oil or the more healthy food consumption or exercise that was the real explanation of lower cholesterol. In other words, the effects of omega-3 fish oil consumption are mixed up with the effects of the other activities that more health conscious people do.

2. **(a)** *P*(type O or Hawaiian-Chinese) = 65,516/145,057 = 0.452 **(b)** *P*(type AB | Hawaiian) = 99/4670 = 0.021 **(c)** *P*(Hawaiian) = 0.032; *P*(Hawaiian | type B) = 0.010. Since these probabilities are not equal, the two events are not independent. **(d)** The probability of randomly selecting a specimen that contains type A blood from the white ethnic group is 50,008/145,057 = 0.345. The probability that at least one of the two samples matches this description = 1 − *P*(neither are type A from white ethnic group) = 1 − $(1 - 0.345)^2$ = 0.571.

3. **(a)** This was an experiment since a treatment was imposed. The researchers added dead cicadas to some plants, while the others served as a control group. **(b)** The distribution of seed mass for the cicada plants has the higher mean. The distribution of seed mass for the cicada plants is slightly skewed to the right, which will pull the mean above its median and toward the higher values. The distribution of seed mass for the control plants is skewed to the left, which will pull the mean of this distribution below its median toward the lower values. Since the medians of both distributions are equal, the mean for the cicada plants lies above the mean for the control plants. **(c)** As stated in part (b), the mean seed mass for the cicada plants is higher than the mean seed mass for the control plants. However, the median seed mass for the two groups is the same. The boxplots show a great deal of overlap between the seed masses of the plants in the two groups. There is some evidence that cicadas can be used effectively as fertilizer, but the difference in the mean seed weights for the two groups isn’t large enough to rule out the chance involved in the random assignment as a plausible explanation.

4. **(a)** Since *P*(diamond) = 1/5, let 8 and 9 represent drawing the card with a diamond and the digits 0 through 7 represent drawing a card without the diamond. Moving left to right across a row, examine one digit at a time. Stop when you get a diamond. Count the number of cards drawn. **(b)** The results: 3, 7, 5, 11, 2, 12, 1, 7, 5, 3 **(c)** Expected number of cards to get the diamond: $\frac{56}{10} = 5.6$. On average,

you will draw 5.6 cards in order to win. **(d)** At $1.00 per card, you will expect to pay $5.60, on average, in order to win $5.00. The game doesn't appear to be fair.

5. **(a)** When comparing data sets it is expected that center, spread, and shape will be addressed. The median of the distribution of cardiovascular fitness measures for Machine A (28) is lower than the median of the distribution of cardiovascular fitness measures for Machine B (38). The range of the distribution of cardiovascular fitness measures for Machine A is substantially smaller than the range of the distribution of cardiovascular fitness measures for Machine B (32 < 57). (Alternatively, the *IQR* for Machine A (37 − 22 = 15) is smaller than the *IQR* for Machine B (47 − 25 = 22).) The distribution of cardiovascular fitness measures for Machine A is reasonably symmetric in shape, while the distribution of cardiovascular fitness measures for Machine B is skewed to the left (toward the lower values). In general, the cardiovascular fitness measures for Machine B tend to be higher than those for Machine A. **(b)** The company should choose Machine B if they want to advertise it as achieving the highest overall gain in cardiovascular fitness. The median for Machine B is higher than it is for Machine A, as is the mean ($\bar{x}_B = 35.4$ versus $\bar{x}_A = 28.9$). **(c)** The company should choose Machine A if they want to advertise it as achieving the most consistent gains in cardiovascular fitness. Machine A exhibits less variation in gains than does Machine B. The *IQR* for Machine A is 15, while the *IQR* for Machine B is 22. Additionally, the standard deviation for Machine A is 9.38, while the standard deviation for Machine B is 16.19. **(d)** Volunteers were used for the experiment and these volunteers may be different in some way from the general population of those who are interested in cardiovascular fitness. Another reason is that the experiment was conducted at only *one* fitness center. Results may vary at other fitness centers in this city and in other cities. If the company had taken a random sample of *all* people who were interested in cardiovascular fitness, it could make inferences about this much larger population. If the company had randomly selected a number of fitness centers from across the country and then randomly selected members of each fitness center to participate in the experiment, the company would be using a stratified random sampling method with the fitness centers serving as the strata.

6. **(a)** Assign each retail sales district a number from 1 to 60 using a random number generator. Order the sales districts numerically. The first 30 are assigned to the monetary incentives group and the remaining 30 to the intangible incentives group. After a specified period of time, compare the mean change in sales for each of the treatment groups to see if there is a difference. **(b)** Assign each retail sales district a number between 01 and 60. Go through the random number table taking two digits at a time. The first 30 two-digit numbers between 01 and 60 to come up are assigned to the monetary incentives group. The remainder will be assigned to the intangible incentives group.
07511 88915 41267 16853 84569 79367 32337 03316
The districts labeled 07, 51, and 18 are the first three to be assigned to the monetary incentives group. **(c)** It would be better to use a matched pairs design. There could be a large variation among the sales figures for the various districts due to the various sizes of those districts across the different regions of the United States. Matching the districts based on their size reduces the effect of variation among the experimental units due to their size on the response variable— sales volume. Pair the two largest districts in size, the next two largest, down to the two smallest districts. For each pair, pick one of the districts and flip a coin. If the flip is "heads," this district is assigned to the monetary incentives group. If it is "tails," this district is assigned to the intangible incentives group. The other district in the pair is assigned to the other group. After a specified period of time, compare the mean change in sales for each of the treatment groups to see if there is a difference.

7. **(a)** Yes. A linear model is reasonable since the scatterplot shows a strong positive linear association between shelf length and weekly sales (in dollars). **(b)** $\hat{y} = 317.94 + 152.68x$, where y = weekly sales (in dollars) and x = shelf length (in feet). **(c)** $\hat{y} = 317.94 + 152.68(5) = 1081.34$. A shelf length of 5 feet would, on average, yield weekly sales of $1081.34. **(d)** The value $s = 22.9212$ represents the standard deviation of the residuals. On average, there is a $22.92 difference between the predicted weekly sales and the actual weekly sales. **(e)** $r^2 = 0.982$. 98.2% of the variation in weekly sales revenue can be explained by the linear regression using shelf length allocated to the house brand as the predictor. **(f)** It would be inappropriate to interpret the intercept, since the data represent sales based on shelf lengths of 3 to 6 feet, and 0 feet falls substantially outside that domain. We would be trying to extrapolate beyond the set of data.