Midterm Review Questions

These questions review Chapters 1-3, to help you prepare for the upcoming midterm. Although this review does not cover every aspect of every topic covered thus far in the semester, it should help you identify some of the areas that you may not fully grasp. Not all questions in this review are multiple choice.

1. Consider the following sample of 4 observations: -2, 8, 52, 107.

a) What is the mean?
b) What is the variance?
c) What is the standard deviation?
d) If we subtracted 57 from every value, what would be the new standard deviation?

2. Consider the following sample data. There are two values that are missing, represented by the question marks.

-2, -5, 17, 12 15, 15, 15, 15, 15, 15, ?, ? ,

What is the value of the median for the entire sample (including the missing values)?

A) 12
B) 15
C) 17
D) 24
E) Impossible to determine from the given information.

3. A study is conducted on second year undergraduate students. Several variables are recorded in the survey. Four of the variables recorded were:

- Brand of car the student owns.
- Amount spent on school supplies in September.
- The time the student waited in line at the bookstore to pay for his/her textbooks.
- Home province of the student.

How many of the variables listed are quantitative?

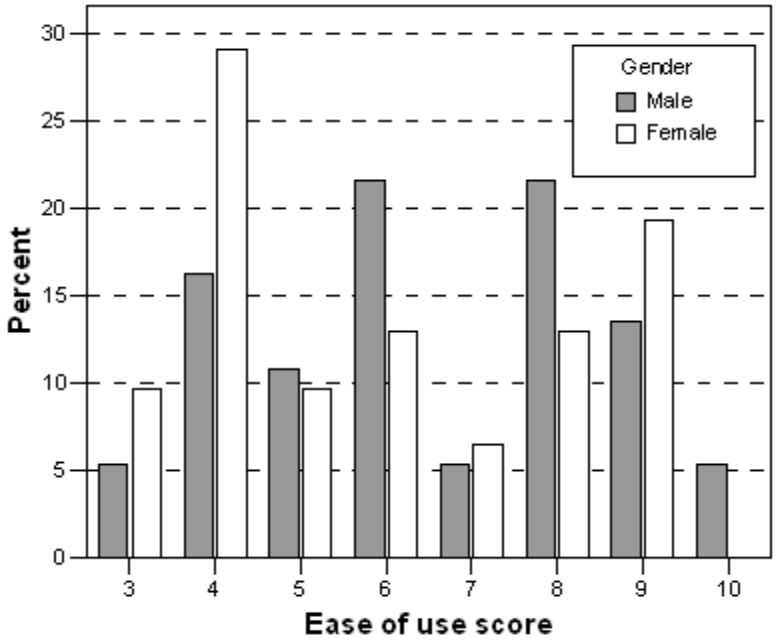
A) 0
B) 1
C) 2
D) 3
E) 4

4. Succinctly state the difference between a parameter and a statistic.

5. Blood clots from inactivity are a concern among frequent airline travelers. A recent study found that out of 8500 frequent airline travelers, 2 developed blood clots. What is the rate of blood clots per 100,000 frequent airline travelers?

A) 4250
B) 235.294
C) 23.529
D) 2.353
E) 0.0002

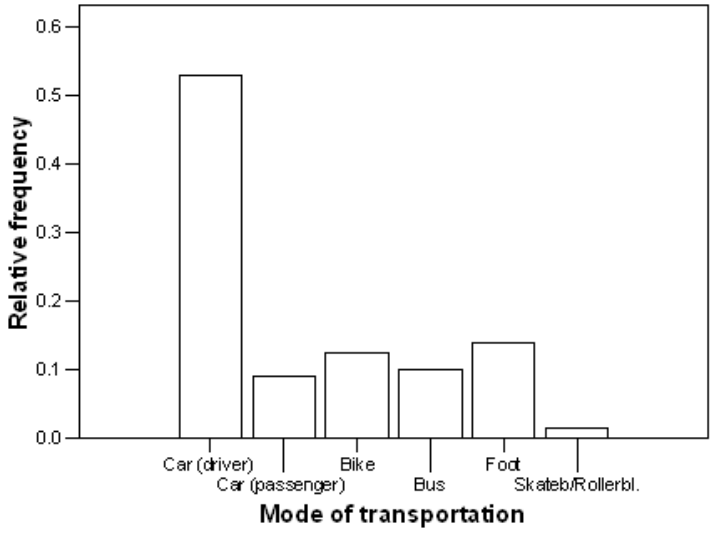
6. A distributor of appliances is doing a customer satisfaction survey for a manufacturer of DVD players. A sample of 68 clients is asked to rate a particular DVD player on appearance, functionality, ease of use, and price on a 1 to 10 scale, where 1 corresponds to the worst rating and 10 to the best rating possible. A bar graph of the ease of use ratings classified by gender is given below:



What percentage of the sampled female clients rated the DVD player as not so easy to use (a rating of 4 or lower)?

A) 18%
B) 29%
C) 38%
D) 49%
E) 62%

The next two questions refer to the following information. A study is being conducted on air quality at a small college in the South. As part of this study, monitors were posted at every entrance to this college from 6 A.M. to 10 P.M. on a randomly chosen day. The monitors recorded the type of transportation used by each person as they entered the campus. Based on the information recorded, the following bar graph was constructed:



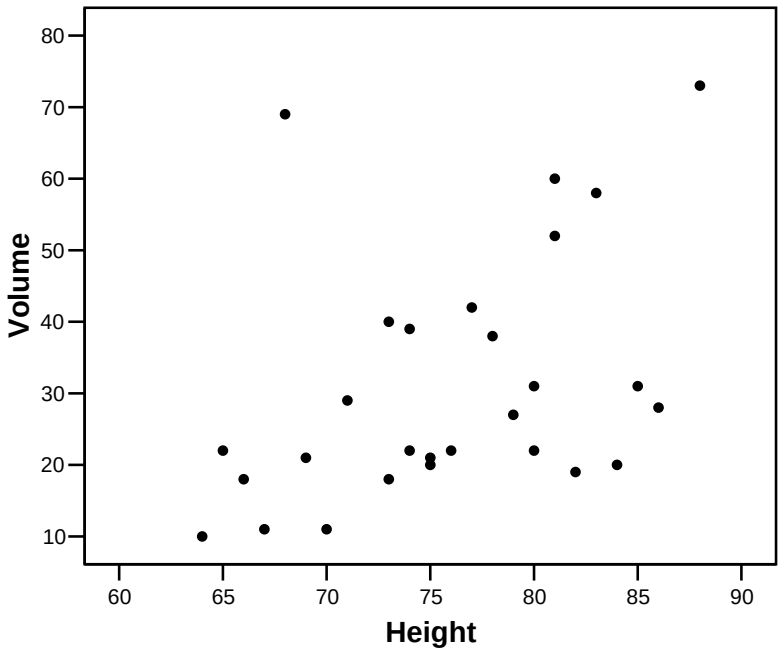
7. Approximately what percentage of people entering campus on this particular day arrived by car?

A) 9%
B) 31%
C) 46%
D) 53%
E) 62%

8. If 1200 people entered campus on this particular day, (approximately) how many people arrived by bus? Choose the response closest to your answer.

A) 10
B) 50
C) 100
D) 120
E) 135

The next two questions refer to the following information. A researcher measured the total height (in feet) and volume of usable lumber (in cubic feet) of each of 32 cherry trees. The goal is to determine if the volume of usable lumber can be estimated from the height of a tree. The results are plotted below:



9. Fill in the blank. The variable _____ is the response variable in this study.

A) Height of the trees.
B) Height of usable lumber.
C) Volume of usable lumber.
D) Number of cherry trees.
E) Cannot tell from information provided.

10. Which of the following descriptions apply to the scatterplot for Volume vs Height?

i) There is a positive association between height and volume.
ii) There is a negative association between height and volume.
iii) There is an outlier in the plot.
iv) The plot is skewed to the left.

A) i) and iii)
B) i) and iv)
C) ii) and iii)
D) ii) and iv)
E) i), iii) and iv)

11. The following R output summarizes the calculations of a regression analysis with $n = 4$.

```
Coefficients:
             Value  Std. Error  t value   Pr(>|t|)
(Intercept) -1.0000   2.1213    -0.4714   0.6838
     x       5.1000   0.7746     6.5840   0.0223
```

a) What is the estimated regression line?
b) Predict the value of y if $x = 2.5$.

The data used to calculate the above R output was:

x	1	2	3	4
y	4.1	8.2	16.3	18.4

c) Calculate the residuals.
d) Compute the correlation between x and y.
e) What fraction of the variation in y can be explained by x?

Use the following to answer questions 12 and 13:

The asking prices (in thousands of dollars) for a sample of 13 houses currently on the market in Neighborville are listed below. For convenience, the data have been ordered.

175 199 205 234 259 275 299 304 317 345 355 384 549

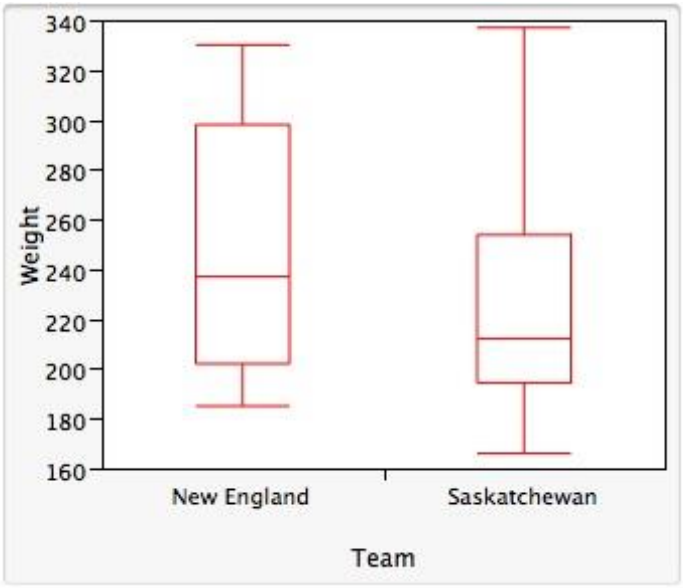
12. What is the five-number summary?

A)	199	234	299	345	549
B)	175	219.5	299	350	549
C)	175	219.5	299	350	384
D)	175	234	299	331	549

13. Use the 1.5 × *IQR* rule to determine if there are any outliers present. What is/are the value(s) of the outlier(s)?

A) No outliers present
B) One outlier: 175
C) One outlier: 549
D) Two outliers: 175 and 549

14. The New England Patriots are a top ranked team in the National Football League (NFL) and the Saskatchewan Roughriders are the 2007 champions of the Canadian Football League (CFL). From the 2007 rosters of these two teams, the weight of the players was determined and the following side-by-side boxplots of their weights is provided below:



Which of the following statements about this side-by-side boxplot is (are) TRUE?

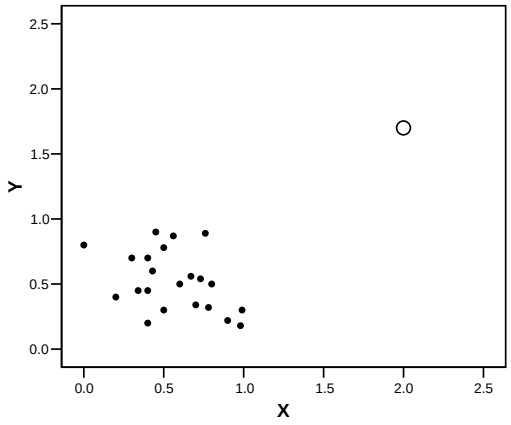
i) The weights of the New England team exhibit less overall variation than the Saskatchewan team.
ii) The median weight for the New England team is higher than the median weight for the Saskatchewan team.
iii) The *IQR* for the Saskatchewan team is greater than the *IQR* of the New England team.

A) i) only.
B) ii) only.
C) iii) only.
D) All of the above.
E) None of the above.

15. Items produced by a manufacturing process are supposed to weigh 90 grams. However, there is variability in the items produced, and they do not all weigh exactly 90 grams. The distribution of weights can be approximated by a Normal distribution with a mean of 90 grams and a standard deviation of 1 gram. What percentage of the items will either weigh less than 87 grams or more than 93 grams?

A) 6%
B) 94%
C) 99.7%
D) 0.3%
E) None of the above.

16. Consider the scatterplot below:



What do we call the point indicated by the plotting symbol O?

A) A residual
B) An influential point
C) A z-score
D) The 3rd quartile
E) A correlation

17. Fill in the blank. If the point indicated by the plotting symbol O were removed from the plot, then the correlation between X and Y would be ____.

A) Close to +1
B) Close to 0
C) Close to –1
D) Higher than the correlation for the original data
E) We cannot tell based on the available information

Use the following to answer questions 18 and 19:

The number of Facebook friends students at a university have are Normally distributed with a mean of 120 and a standard deviation of 20.

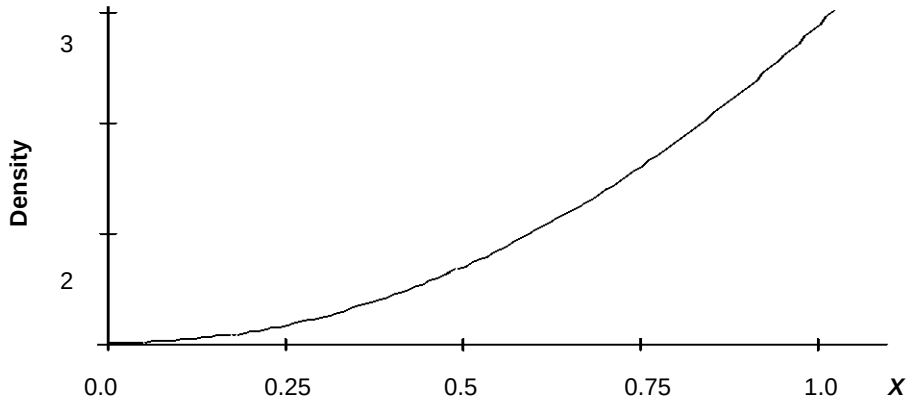
18. What percent of students have at least 100 Facebook friends?

A) 8.24%
B) 15.86%
C) 42.07%
D) 84.13%
E) None of the above.

19. What percent of students do we expect to have exactly 1000 Facebook friends?

A) 8.24%
B) 15.86%
C) 42.07%
D) 84.13%
E) None of the above.

20. For the density curve below, which of the following is true?



A) The mean and median are equal
B) The mean is greater than the median.
C) The mean is less than the median.
D) The mean could be either greater than or less than the median, but not equal.
E) Not enough information is provided to answer the question.

21. Which of the following statements about a density curve are FALSE?

A) A density curve always has area beneath it equal to 1.
B) A density curve can adequately describe outliers observed in data.
C) A density curve is always on or above the horizontal axis.
D) A density curve comes in many shapes, some of which are symmetric while others are skewed.
E) The area under a density curve above any range of values is the proportion of all observations that fall in that range.

22. The time to complete an exam is approximately Normal with a mean of 70 minutes and a standard deviation of 10 minutes. Using the 68-95-99.7 rule, what percentage of students will complete the exam in under an hour?

A) 68%
B) 32%
C) 16%
D) 5%
E) 1%

23. Which of the following statements about Normal quantile plots is (are) FALSE?

A) In constructing a Normal quantile plot, each data point is plotted on one axis and the corresponding Normal scores on the other axis.
B) The Normal quantile plot is a very useful graphical tool for assessing the adequacy of the Normal model.
C) If the points on a Normal quantile plot lie close to a straight line, the plot indicates that the Normal model is an adequate representation for the data.
D) Because you will see the usual mound-like appearance of the Normal distribution on a histogram, it is more helpful than the quantile plot for assessing Normality.
E) On a quantile plot, outliers will appear as points that are far away from the overall pattern of the plot.

24. Which of the following statements is (are) TRUE?

A) A two-way table is a useful way to summarize data when two categorical variables are measured on the same individuals or cases.
B) Simpson's paradox is an example of the potential effect of a lurking variable on an observed association between two categorical variables.
C) If the counts in each cell of a two-way table are divided by the total number of observations, the result is the joint distribution of the two categorical variables.
D) All of the above are true.
E) Only A and C are true.

Use the following data to answer questions 25 to 27:

A survey was conducted involving 303 subjects concerning their preferences with respect to the size of car they would consider purchasing. The following table shows the count of the responses by gender of the respondents:

	Preferred Size of Car			
Gender	Small	Medium	Large	Total
Female	58	63	17	138
Male	79	61	25	165
Total	137	124	42	303

25. The data are to be summarized by constructing marginal distributions. In the marginal distribution for car size, the entry for Medium is _____.

A) 0.370
B) 0.409
C) 0.457
D) 0.508
E) None of the above.

26. In the conditional distribution for preference of car size among male respondents, the entry for Large cars is _____.

A) 0.056
B) 0.139
C) 0.152
D) 0.405
E) None of the above.

27. Across all respondents, the proportion of female respondents who preferred small cars is _____.

A) 0.191
B) 0.418
C) 0.423
D) 0.452
E) None of the above.

28. The California Department of State Police keeps track of the number of points received for various traffic violations by drivers. The department is interested in examining the relationship between the number of points received and the insurance premium. Some information on the point category and the insurance premium category is given below:

Insurance premium category	Traffic Violation Point category		
	Low	Medium	High
Cheap	.12	.38	.50
Medium	.29	.33	.27
Expensive	.59	.29	.23

Which distribution is displayed in the above table?

A) The joint distribution of premium category and point category.
B) The marginal distribution of point category.
C) The conditional distribution of premium category given point category.
D) The conditional distribution of point category given premium category.
E) The table does not display any distribution.

29. A study of the salaries of full professors at a small university shows that the median salary for female professors is considerably less than the median male salary. Further investigation shows that the median salaries for male and female full professors are about the same in every department (English, Physics, etc.) of the university. Which phenomenon explains the reversal in this example?

A) Extrapolation
B) Simpson's paradox
C) Causation
D) Correlation
E) None of the above

Use the following to answer questions 30–32:

Prior to graduation, a high school class was surveyed about their plans after high school. The table below displays the results by gender:

	Plans				
Gender	4-yr college	2-yr college	Military	Work	Other
Male	198	36	4	14	16
Female	176	36	1	3	5

30. If the data are going to be summarized by computing the marginal distribution of plans after high school, what would be the entry for 4-year college?

A) 0.529
B) 0.739
C) 0.765
D) 374
E) None of the above.

31. If the data are going to be summarized by computing the conditional distributions of plans after high school for male and female high school students, respectively, what would be the entry for "male" and "2-year college"?

A) 0.074
B) 0.134
C) 0.5
D) 39.46
E) None of the above.

32. If the data are going to be summarized by computing the conditional distributions of gender given plans after high school, what would be the entry for "male" and "2-year college"?

A) 0.074
B) 0.134
C) 0.36
D) 0.5
E) None of the above.

33. A researcher notices in a random sample of adults that those who took larger doses of ginko-ginseng supplements tended to stay alert for longer periods of time than those who took smaller amounts. He further noticed that those on the larger doses of ginko-ginseng were more likely to drink more coffee. The researcher is interested in assessing the effect of ginko-ginseng on brain-stimulation. The amount of coffee consumed is what type of variable in this study?

(A) skewed
(B) response
(C) treatment
(D) confounder
(E) exposure

The next three questions refer to the following information: An investigator is recruiting subjects from southwestern Ontario through radio advertisements for a double blind randomized study to test the effectiveness of a new experimental drug on treating symptoms of dust allergies compared to a gold standard drug. Subjects will be expected to participate in the study over a span of 2 years. To be eligible in the study subjects must have dust allergies, and as an incentive to participate, subjects are guaranteed $200 for every year of completion.

34. State what the population, sample, treatment, and response are.

35. How many of the following statements are TRUE?

(i) Subjects are randomly selected.
(ii) This study is an experiment.
(iii) Results from this study are generalizable to all sufferers of dust allergies in southwestern Ontario.
(iv) The doctor who will assess the extent of allergy symptoms is unaware of which drug the subject is randomized to.

36. Assume 12 subjects have been recruited into the study. Use Table B, the random digit table at the back of your book, to assign treatment to half of the subjects. Label your subjects as 01, 02, ..., 12 and state who receives treatment. Start at Line 198 of the table.

37. When do we use a block design instead of a completely randomized design? What is the main advantage of using a block design?

38. Which one of the following statements is true?

(A) A matched-pairs design is a specific type of completely randomized design.
(B) Randomizing treatment enables us to use far fewer experimental units to achieve the same results.
(C) The standard deviation is always larger than the interquartile range, as long as all values of the data set are positive.
(D) The median is always less than the standard deviation if all values of the data set are negative.
(E) The median is a measure of the variability of a data set.

39. Which of the following 4 statements are reasons for using a double-blind experiment? There may be more than one correct answer.

(i) To reduce possible bias in the person taking the measurements.
(ii) To counter the placebo effect.
(iii) To reduce the number of subjects needed in an experiment.
(iv) To enable a block design to be used.

40. A daycare provider notices that her colleagues periodically ask the toddlers if they need to use the washroom before taking them to the washroom. She then reads a blog on a parenting discussion forum in which a parent also asked her toddler if he wanted to use the washroom before taking him to the washroom, and that her son was potty trained in 2 weeks. Based on this information, the daycare provider starts asking the toddlers in her class if they need to use the toilet before taking them to the washroom. What type of study is her decision based on?

(A) a double blind experiment
(B) anecdotal evidence
(C) observational study based on a random sample
(D) matched pairs experiment
(E) observational study based on volunteers

27. The method of least squares obtains an estimated regression equation that minimizes:
(A) the sum of the vertical distances between the points and the estimated regression equation.
(B) the sum of the perpendicular distances between the points and the estimated regression equation.
(C) the sum of the squared vertical distances between the points and the estimated regression equation.
(D) the sum of the squared perpendicular distances between the points and the estimated regression equation.
(E) the sum of the squared horizontal distances between the points and the estimated regression equation.

32. The type of graph used in linear regression and correlation analysis to illustrate the bivariate data is called a:
(A) box plot.
(B) relative frequency histogram.
(C) pie chart.
(D) cumulative frequency polygon.
(E) scatter diagram.

33. In simple linear regression, the units of the intercept must be:
(A) the same as the units of the dependent variable.
(B) the same as the units of the independent variable.
(C) the same as the units of the slope.
(D) the units of dependent variable divided by the units of the independent variable.
(E) dimensionless (has no units).