INSTRUCTOR'S SOLUTIONS MANUAL

JAMES LAPP

STATISTICAL METHODS FOR THE SOCIAL SCIENCES

FIFTH EDITION

Alan Agresti

University of Florida

Reproduced by Pearson from electronic files supplied by the author.

Publishing as Pearson, 330 Hudson Street, NY NY 10013

ISBN-13: 978-0-13-451277-8
ISBN-10: 0-13-451277-4

CONTENTS

Chapter 1: Introduction 1

Chapter 2: Sampling and Measurement 3

Chapter 3: Descriptive Statistics 7

Chapter 4: Probability Distributions 21

Chapter 5: Statistical Inference: Estimation 29

Chapter 6: Statistical Inference: Significance Tests 37

Chapter 7: Comparison of Two Groups 47

Chapter 8: Analyzing Association Between Categorical Variables 59

Chapter 9: Linear Regression and Correlation 67

Chapter 10: Introduction to Multivariate Relationships 83

Chapter 11: Multiple Regression and Correlation 89

Chapter 12: Regression with Categorical Predictors: Analysis of Variance Methods 103

Chapter 13: Multiple Regression with Quantitative and Categorical Predictors 111

Chapter 14: Model Building with Multiple Regression 117

Chapter 15: Logistic Regression: Modeling Categorical Responses 127

Chapter 16: An Introduction to Advanced Methodology 135

Chapter 1: Introduction

1.1. (a) an individual Prius (automobile)

(b) All Prius automobiles used in the EPA tests.

(c) All Prius automobiles that are or may be manufactured.

1.2. (a) all 7.3 million voters is the population. The sample is the 1824 voters surveyed.

(b) A statistic is the 60.5% who voted for Brown from the exit poll sample of size 1824; a parameter is the 60.0% who actually voted for Brown.

1.3. (a) all students at the University of Wisconsin

(b) A statistic, since it's calculated only for the 100 sampled students.

1.4. The values are statistics, since they is based on the 1028 adults in the sample.

1.5. (a) all adult Americans

(b) Proportion of all adult Americans who would answer definitely or probably true.

(c) The sample proportion 0.523 estimates the population proportion.

(d) No, it is a prediction of the population value but will not equal it exactly, because the sample is only a very small subset of the population.

1.6. (a) The most common response was 2 hours per day.

(b) This is a descriptive statistic because it describes the results of a sample.

1.7. (a) A total of 85.7% said “yes, definitely” or “yes, probably.”

(b) In 1998, a total of 85.8% said “yes, definitely” or “yes, probably.”

(c) A total of 74.4% said “yes, definitely” or “yes, probably.” The percentages of yes responses were higher for HEAVEN than for HELL.

1.8. (a) Statistics, since they're based on a sample of 60,000 households, rather than all households.

(b) Inferential, predicting for a population using sample information.

1.9. The correct answer is (a).

1.10.

Race	Age	Sentence	Felony?	Prior Arrests	Prior Convictions
white	19	2	no	2	1
black	23	1	no	0	0
white	38	10	yes	8	3
Hispanic	20	2	no	1	1
white	41	5	yes	5	4

1.11. (a) There are 60 rows in the data.

(b) Answers will vary.

1.12. Answers will vary.

1.13. Answers will vary.

1.14. (a) A statistic is a numerical summary of the sample data, while a parameter is a numerical summary of the population. For example, consider an exit poll of voters on election day. The proportion voting for a particular candidate is a statistic. Once all of the votes have been counted, the proportion of voters who voted for that candidate would be known (and is the parameter).

(b) Description deals with describing the available data (sample or population), whereas inference deals with making predictions about a population using information in the sample. For example, consider a sample of voters on election day. One could use descriptive statistics to describe the voters in terms of gender, race, party, etc., and inferential statistics to predict the winner of the election.

1.15. If you have a census, you do not need to use the information from a sample to describe the population since you have information from the population as a whole.

1.16. (a) The descriptive part of this example is that the average age in the sample is 24.1 years.

(b) The inferential part of this example is that the sociologist estimates the average age of brides at marriage for the population to between 23.5 and 24.7 years.

(c) The population of interest is women in New England in the early eighteenth century.

1.17. (a) A statistic is the 78% of the sample of subjects interviewed in the UK who said yes.

(b) A parameter is the true percent of the 50 million adults in the UK who would say yes.

(c) A descriptive analysis is that the percentage of yes responses in the survey varied from 56% (in Denmark) to 95% (in Cyprus).

(d) An inferential analysis is that the percentage of adults in the UK who would say yes falls between 75% and 81%.

1.18. Answers will vary.

1.19. Answers will vary.

Chapter 2: Sampling and Measurement

2.1. (a) Discrete variables take a finite set of values (or possible all nonnegative integers), and we can enumerate them all. Continuous variables take an infinite continuum of values.

(b) Categorical variables have a scale that is a set of categories; for quantitative variables, the measurement scale has numerical values that represent different magnitudes of the variable.

(c) Nominal variables have a scale of unordered categories, whereas ordinal variables have a scale of ordered categories. The distinctions among types of variables are important in determining the appropriate descriptive and inferential procedures for a statistical analysis.

2.2. (a) quantitative
(b) categorical
(c) categorical
(d) quantitative
(e) categorical
(f) quantitative
(g) categorical
(h) quantitative
(i) categorical

2.3. (a) ordinal
(b) nominal
(c) interval
(d) nominal
(e) nominal
(f) ordinal
(g) interval
(h) ordinal
(i) nominal
(j) interval
(k) ordinal

2.4. (a) nominal
(b) nominal
(c) ordinal
(d) interval
(e) interval
(f) interval
(g) ordinal
(h) interval
(i) nominal
(j) interval

2.5. (a) interval
(b) ordinal
(c) nominal

2.6. (a) state of residence
(b) number of siblings
(c) social class (high, medium, low)
(d) student status (full time, part time)
(e) Number of cars owned.
(f) Time (in minutes) needed to complete an exam.
(g) number of siblings

2.7. (a) Ordinal, since there is a sense of order to the categories.
(b) discrete
(c) These values are statistics, because they apply to a sample of size 1962, not the entire population.

2.8. ordinal

2.9. The correct responses are (b), (c), (d), (e) and (f).

2.10. The correct responses are (a), (c), (e), and (f).

2.11. Answers will vary.

2.12. Number names 00001 to 52000. Answers will vary 6907.

2.13. (a) observational study

(b) experiment

(c) observational study

(d) experiment

2.14. (a) Experimental study, since the researchers are assigning subjects to treatments.

(b) An observational study could look those who grew up in nonsmoking or smoking environments and examine incidence of lung cancer.

2.15 (a) Sample-to-sample variability causes the results to vary.

(b) The sampling error for the Gallup poll is –2.1% for Obama and 2.8% for Romney.

2.16. (a) This is a volunteer sample because viewers chose whether to call in.

(b) The mail-in questionnaire is a volunteer sample because readers chose whether to respond.

2.17. The first question is confusing in its wording. The second question has clearer wording.

2.18. (a) Skip number is $k = 52,000/5 = 10,400$. Randomly select one of the first 10,400 names and then skip 10,400 names to get each of the next names. For example, if the first name picked is 01536, the other four names are $01536 + 10400 = 11936,\ 11936 + 10400 = 22336,\ 22336 + 10400 = 32736,\ 32736 + 10400 = 43136$.

(b) We could treat the pages as clusters. We would select a random sample of pages, and then sample every name on the pages selected. Its advantage is that it is much easier to select the sample than it is with random sampling. A disadvantage is as follows: Suppose there are 100 "Martinez" listings in the directory, all falling on the same page. Then with cluster sampling, either all or none of the Martinez families would end up in the sample. If they are all sampled, certain traits which they might have in common (perhaps, e.g., religious affiliation) might be over-represented in the sample.

2.19. Draw a systematic sample form the student directory, using skip number $k = 5000/100 = 50$.

2.20. (a) This is not a simple random sample since the sample will necessarily have 25 blacks and 25whites. A simple random sample may or may not have exactly 25 blacks and 25 whites.

(b) This is stratified random sampling. You ensure that neither blacks nor whites are over-sampled.

2.21. (a) the clusters

(b) The subjects within every stratum.

(c) The main difference is that a stratified random sample uses every stratum, and we want to compare the strata. By contrast, we have a sample of clusters, and not all clusters are represented—the goal is not to compare the clusters but to use them to obtain a sample.

2.22. (a) Categorical are GE, VE, AB, PI, PA, RE, LD, AA; quantitative are AG, HI, CO, DH, DR, NE, TV, SP, AH.

(b) Nominal are GE, VE, AB, PA, LD, AA; ordinal are PI and RE; interval are AG, HI, CO, DH, DR, NE, TV, SP, AH.

2.23. Answers will vary.

2.24. (a) Draw a systematic sample from the student directory, using skip number $k = N/100$, where $N =$ number of students on the campus.

(b) High school GPA on a 4-point scale, treated as quantitative, interval, continuous; math and verbal SAT on a 200 to 800 scale, treated as quantitative, interval, continuous; whether work to support study (yes, no), treated as categorical, nominal, discrete; time spent studying in average day, on scale (none, less than 2 hours, 2–4 hours, more than 4 hours), treated as quantitative, ordinal, discrete.

2.25. This is nonprobability sampling; certain segments may be over- or under-represented, depending on where the interviewer stands, time of day, etc. Quota sampling fails to incorporate randomization into the selection method.

2.26. Responses can be highly dependent on nonsampling errors such as question wording.

2.27. (a) This is a volunteer sample, so results are unreliable; e.g., there is no way of judging how close 93% is to the actual population who believe that benefits should be reduced.

(b) This is a volunteer sample; perhaps an organization opposing gun control laws has encouraged members to send letters, resulting in a distorted picture for the congresswoman. The results are completely unreliable as a guide to views of the overall population. She should take a probability sample of her constituents to get a less biased reaction to the issue.

(c) The physical science majors who take the course might tend to be different from the entire population of physical science majors (perhaps more liberal minded on sexual attitudes, for example). Thus, it would be better to take random samples of students of the two majors from the population of all social science majors and all physical science majors at the college.

(d) There would probably be a tendency for students within a given class to be more similar than students in the school as a whole. For example, if the chosen first period class consists of college-bound seniors, the members of the class will probably tend to be less opposed to the test than would be a class of lower achievement students planning to terminate their studies with high school. The design could be improved by taking a simple random sample of students, or a larger random sample of classes with a random sample of students then being selected from each of those classes (a two-stage random sample).

2.28. A systematic sample with a skip number of 7 (or a multiple of 7) would be problematic since the sampled editions would all be from the same day of the week (e.g., Friday). The day of the week may be related to the percentage of newspaper space devoted to news about entertainment.

2.29. Because of skipping names, two subjects listed next to each other on the list cannot both be in the sample, so not all samples are equally likely.

2.30. If we do not take a disproportional stratified random sample, we might not have enough Native Americans in our sample to compare their views to those of other Americans.

2.31. If a subject is in one of the clusters that is not chosen, then this subject can never be in the sample. Not all samples are equally likely.

2.32. Answers will vary

2.33. The nursing homes can be regarded as clusters. A systematic random sample is taken of the clusters, and then a simple random sample is taken of residents from within the selected clusters.

2.34. The best answer is (b).

2.35. The best answer is (c).

2.36. The best answer is (c).

2.37. The best answer is (a).

2.38. False; this is a convenience sample.

2.39. False; this is a voluntary response sample.

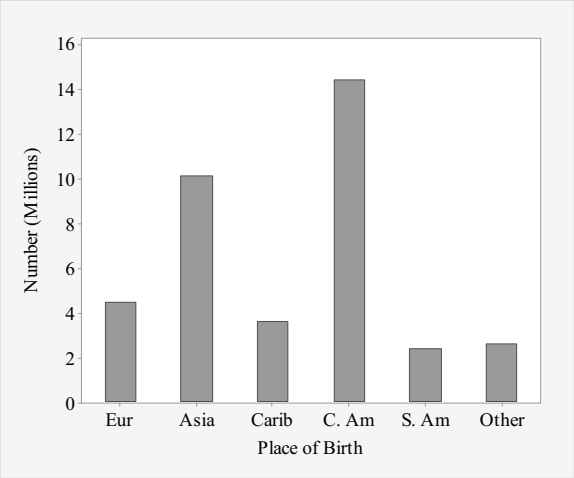
2.40. An annual income of $40,000 is twice the annual income of $20,000. However, 70 degrees Fahrenheit is not twice as hot as 35 degrees Fahrenheit. (Note that income has a meaningful zero and temperature does not.) IQ is not a ratio-scale variable.

Chapter 3: Descriptive Statistics

3.1. (a)

Place of Birth	Number (Millions)	Relative Frequency
Europe	4.5	4.5 / 37.6 = 12.0%
Asia	10.1	10.1 / 37.6 = 26.9 %
Caribbean	3.6	3.6 / 37.6 = 9.6%
Central America	14.4	14.4 / 37.6 = 38.3%
South America	2.4	2.4 / 37.6 = 6.4%
Other	2.6	2.6 / 37.6 = 6.9%
Total	37.6	

(b)



(c) "Place of birth" is categorical.

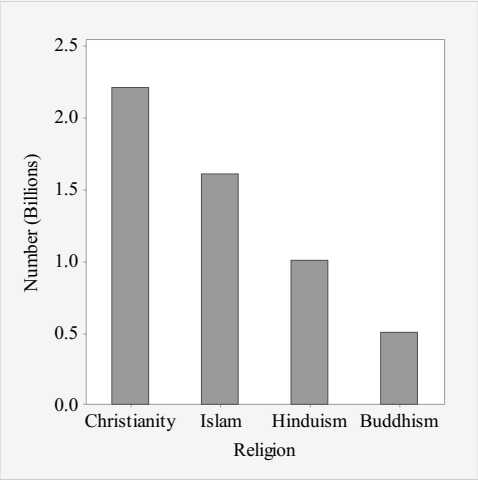
(d) The mode is Central America.

3.2. (a)

Religion	Number (Billions)	Relative Frequency
Christianity	2.2	2.2 / 5.3 = 41.5%
Islam	1.6	1.6 / 5.3 = 30.2%
Hinduism	1.0	1.0 / 5.3 = 18.9%
Buddhism	0.5	0.5 / 5.3 = 9.4%
Total	5.3	

3.2 (continued)

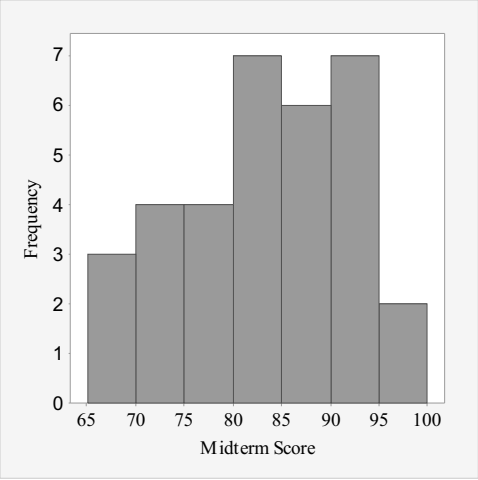
(b)



(c) The mean or median cannot be calculated for these data since they are categorical. The mode of these four religions is Christianity. Christianity is also the mode of all religions.

3.3. (a) There are 33 students. The minimum score is 65, and the maximum score is 98.

(b)

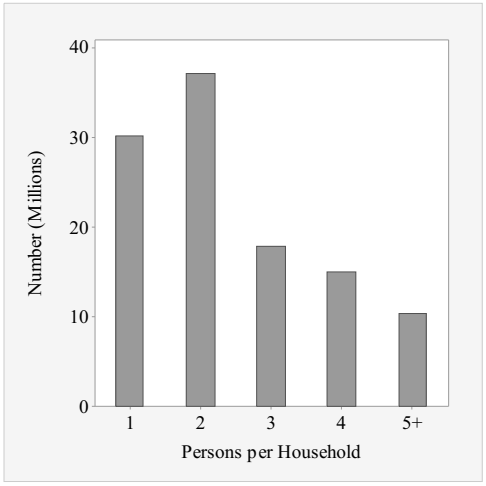


3.4. (a)

Persons per Household	Number (Millions)	Relative Frequency
1	30.1	30.1/110.4 = 27.3%
2	37.1	37.1/110.4 = 33.6%
3	17.8	17.8/110.4 = 16.1%
4	15.0	15.0/110.4 = 13.6%
5 or more	10.4	10.4/110.4 = 9.4%
Total	110.4	

3.4 (continued)

(b) The distribution is right skewed.

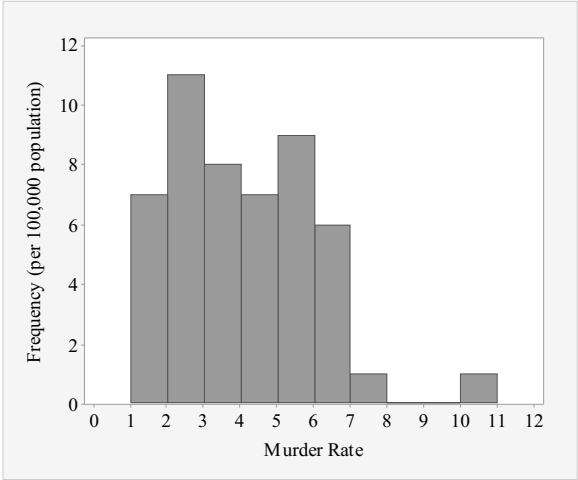


(c) The median household size is 2 persons, and the mode is also 2 persons.

3.5. (a)

Murder Rate	Frequency	Relative Frequency
1+	7	14%
2+	11	22%
3+	8	16%
4+	7	14%
5+	9	18%
6+	6	12%
7+	1	2%
8+	0	0%
9+	0	0%
10+	1	2%
Total	50	

(b)



The distribution appears to be somewhat skewed right, with outlier at 10.8.

3.5 (continued)

(c)

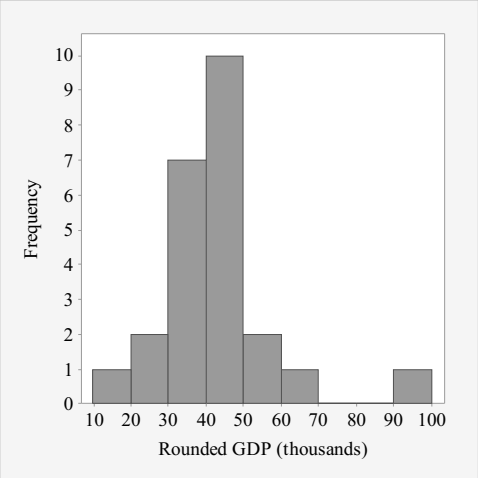
Stem (1)	Leaves (0.1)
1	4567778
2	00122344899
3	13348899
4	2356678
5	001444568
6	012445
7	2
8	
9	
10	8

The stem-and-leaf plot looks like the histogram turned on its side.

3.6. (a) GDP is rounded to the nearest thousand

Stem (10,000s)	Leaves (1,000s)
1	9
2	66
3	2456678
4	0033333444
5	34
6	5
7	
8	
9	1

(b)



(c) The outlier in each plot is Luxembourg.

3.7. (a) The mean is 129.5 and the median is 102

(b) The mean would be higher because of the skew to the right and the extreme outlier (716 for the U.S.)

(c) Without the United States, the mean is 104 and the median is 98. There is a greater effect on mean.

3.8. (a) The mean is $(0.4+2.2+6.2+1.7+1.8+0.9+12.2+17.6)/8 = 5.375$ metric tons per person. The median is 2.0 metric tons per person.

(b) The mean is now $(0.4+2.2+6.2+1.7+1.8+0.9+12.2+17.6+40.3)/9 = 9.26$ metric tons per person. The median is 2.2 metric tons per person. Qatar had a greater impact on the mean.

3.9. (a) The response "not far enough" is the mode.

(b) The median is "not far enough." We cannot the mean with these data since they are categorical.

3.10. (a) For the data from the previous 2, the mean is 16.6 days, the median is 12 days, and the standard deviation is 13.9. For the data from 25 years ago, the mean was 27.6 days, the median was 24 days, and the standard deviation is 12.4. The mean has decreased by 11 days, and the median has decreased by 12 days since 25 years ago. The variability in length of stay was slightly lower 25 years ago.

(b) Of the 11 observations, the median is 13 days. We cannot calculate the mean, but substituting 40 for the censored observation gives a mean of 18.7 days.

3.11. (a)

TV Hours	Frequency	Relative Frequency
0	120	7.2
1	310	18.6
2	437	26.2
3	293	17.6
4	227	13.6
5	113	6.8
6	75	4.5
7+	94	5.6
Total	1669	100.1

(b) The distribution is unimodal and right skewed.

(c) The median is the 835th data value, which is 2.

(d) The mean is larger than 2 because the data is skew right by a few high values.

3.12.

Eastern Europe		Middle East
Leaves (1)	Stem (10)	Leaves (1)
	0	4
	1	9
	2	9
	3	489
	4	029
	5	
87	6	
995554	7	
7321100	8	1

Eastern Europe: mean = 77.7, standard deviation = 5.4, minimum = 67.0, Q1 = 75.0, median = 79.0, Q3 = 81.0, maximum = 87.0

Middle East: mean = 37.5, standard deviation = 20.0, minimum = 4.0, Q1 = 29.0, median = 38.5, Q3 = 42.0, maximum = 81.0

Female economic activity seems greater, on average, in Eastern Europe than in the Middle East. Except for standard deviation, all of the descriptive statistics in Eastern Europe exceed the descriptive statistics in the Middle East. There appear to be more women in the labor force (per 100 men) in Eastern Europe than in the Middle East.