Complete Solutions Manual to Accompany

An Introduction to Statistical Methods and Data Analysis

SEVENTH EDITION

R. Lyman Ott

Michael Longnecker
Texas A&M University
College Station, TX

Prepared by

John Daniel Draper
The Ohio State University, Columbus, OH

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

ISBN-13: 978-1-305-26990-3
ISBN-10: 1-305-26990-X

Cengage Learning
20 Channel Center Street, Floor 4
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at: **www.cengage.com/global**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**.

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**.

Printed in the United States of America
1 2 3 4 5 6 7 17 16 15 14 13

Contents

Chapter 1

Statistics and the Scientific Method

1.1

a. The population of interest is all salmon released from fish farms located in Norway.
b. The samples are the two batches of salmon released (1,996 and 2,499 in northern and southern Norway, respectively).
c. The migration pattern and survival of salmon released from fish farms.
d. Since the sample is only a small proportion of the whole population, it is necessary to evaluate what the mean weight may be for any other random selection of farmed salmon.

1.2

a. All private water wells.
b. The 100 private water wells in or near the Barnett Shale in Texas.
c. The level of contaminants in the water wells.
d. We want to relate the level of contaminants of the 100 points in the sample to the level in the whole suspect area. Thus we need to know how accurate a portrayal of the population is provided by the 100 points in the sample.

1.3

a. All families that have had option of SNAP (food stamps).
b. 60,782 examined over the time period of 1968 to 2009.
c. Adult health and economic outcomes (specifically, the incidence of metabolic health outcomes and economic self-sufficiency).
d. In order to evaluate how closely the sample families represent the American population over this time period.

1.4

a. All head impacts resulting from playing football over a given period of time.
b. The 1,281,444 head impacts recorded.
c. The number (or percent) of concussions suffered through these impacts.
d. The advances in tackling techniques imply that there is variability in how a tackle is performed. We need to see if our sample was representative of the hits that may be sustained.

1.5

a. The population of interest is the population of those who would vote in the 2004 senatorial campaign.
b. The population from which the sample was selected is registered voters in this state.
c. The sample will adequately represent the population, unless there is a difference between registered voters in the state and those who would vote in the 2004 senatorial campaign.
d. The results from a second random sample of 5,000 registered voters will not be exactly the same as the results from the initial sample. Results vary from sample to sample. With either sample we hope that the results will be close to that of the views of the population of interest.

1.6

a. The professor's population of interest is college freshmen at his university.
b. The sampled population is all freshmen enrolled in HIST 101.
c. Yes, there is a major difference in the two populations. Those enrolled in HIST 101 may not accurately reflect the population of all freshmen at his university. For example, they might be more interested in history.
d. Had the professor lectured on the American Revolution, those students in HIST 101 would be more likely to know which country controlled the original 13 states prior to the American Revolution than other freshmen at the university.

Chapter 2

Using Surveys and Experimental Studies to Gather Data

2.1

a. The explanatory variable is level of alcohol drinking. One possible confounding variable is smoking. Perhaps those who drink more often also tend to smoke more, which would impact incidence of lung cancer. To eliminate the effect of smoking, we could block the experiment into groups (e.g., nonsmokers, light smokers, heavy smokers).
b. The explanatory variable is obesity. Two confounding variables are hypertension and diabetes. Both hypertension and diabetes contribute to coronary problems. To eliminate the effect of these two confounding variables, we could block the experiment into four groups (e.g., hypertension and diabetes, hypertension but no diabetes, diabetes but no hypertension, neither hypertension nor diabetes).

2.2

a. The explanatory variable is the new blood clot medication. The confounding variable is the year in which patients were admitted to the hospital. Because those admitted to the hospital the previous year were not given the new blood clot medication, we cannot be sure that the medication is working or if something else is going on. We can eliminate the effects of this confounding by randomly assigning stroke patients to the new blood clot medication or a placebo.
b. The explanatory variable is the software program. The confounding variable is whether students choose to stay after school for an hour to use the software on the school's computers. Those students who choose to stay after school to use the software on the school's computers may differ in some way from those students who do not choose to do so, and that difference may relate to their mathematical abilities. To eliminate the effect of the confounding variable, we could randomly assign some students to use the software on the school's computers during class time and the rest to stay in class and learn in a more traditional way.

2.3 Possible confounding factors include student-teacher ratios, expenditures per pupil, previous mathematics preparation, and access to technology in the inner city schools. Adding advanced mathematics courses to inner city schools will not solve the discrepancy between minority students and white students, since there are other factors at work.

2.4 There may be a difference in student-teacher ratios, expenditures per pupil, and previous preparation between the schools that have a foreign language requirement and schools that do not have a foreign language requirement.

2.5 The relative merits of the different types of sampling units depends on the availability of a sampling frame for individuals, the desired precision of the estimates from the sample to the population, and the budgetary and time constraints of the project.

2.6 She could conduct a stratified random sample in which the states serve as the stratum. A simple random sample could then be selected within each state. This would provide information concerning the differences between the states along with the individual opinions of the employees.

2.7

a. All residents in the county.
b. All registered voters.
c. Survey nonresponse – those who responded were probably the people with much stronger opinions than those who did not respond, which then makes the responses not representative of the responses of the entire population.

2.8

a. In the first scenario, people would be more willing to lie about using a biodegradable detergent because there is no follow up to verify and individuals usually prefer to appear environmentally conscious. The second survey has a check in place to verify the answers given are truthful.
b. The first survey would likely yield a higher percentage of those who say they use a biodegradable detergent. The second may anger the individuals who tell the truth as if their honesty is being tested.

2.9

a. Alumni (men only?) who graduated from Yale in 1924.
b. No. Alumni whose addresses were on file 25 years later would not necessarily be representative of their class.
c. Alumni who responded to the mail survey would not necessarily be representative of those who were sent the questionnaires. Income figures may not be reported accurately (intentionally), or may be rounded off to the nearest $5,000, say, in a self-administered questionnaire.
d. Rounding income responses would make the figure $25,111 unlikely. The fact that higher income respondents would be more likely to respond (bragging), and the fact that incomes are likely to be exaggerated, would tend to make the estimate too high.

2.10

a. Simple random sampling.
b. Stratified sampling.
c. Cluster sampling.

2.11

a. Simple random sampling.
b. Stratified sampling.
c. Cluster sampling.

2.12

a. Stratified sampling. Stratify by job category and then take a random sample within each job category. Different job categories will use software applications differently, so this sampling strategy will allow us to investigate that.
b. Systematic random sampling. Sample every tenth patient (starting from a randomly selected patient from the first ten patients). Provided that there is no relationship between the type of patient and the order that the patients come into the emergency room, this will give us a representative sample.

2.13

a. Stratified sampling. We should stratify by type of degree and then sample 5% of the alumni within each degree type. This method will allow us to examine the employment status for each degree type and compare among them.

b. Simple random sampling. Once we find 100 containers we will stop. Still it will be difficult to get a completely random sample. However, since we don't know the locations of the containers, it would be difficult to use either a stratified or cluster sample.

2.14

a. Water temperature and Type of hardener
b. Water temperature: 175 °F and 200 °F; Type of hardener: H_1, H_2, H_3
c. Manufacturing plants
d. Plastic pipe
e. Location on Plastic pipe
f. 2 pipes per treatment
g. Covariates: None
h. 6 treatments: (175 °F, H_1), (175 °F, H_2), (175 °F, H_3), (200 °F, H_1), (200 °F, H_2), (200 °F, H_3)

2.15

This is an example where there are two levels of Experimental units, and the analysis is discussed in Chapter 18.

To study the effect of month:

a. Factors: Month
b. Factor levels: 8 levels of month (Oct - May)
c. Block = each section
d. Experimental unit (Whole plot EU) = each tree
e. Measurement unit = each orange
f. Replications = 8 replications of each month
g. Covariates = none
h. Treatments = 8 treatments (Oct – May)

To study the effect of location:

a. Factors: Location
b. Factor levels: 3 levels of location (top, middle, bottom)
c. Block = each section
d. Experimental unit (Split plot EU) = each location tree
e. Measurement unit = each orange
f. Replications = 8 replications of each location
g. Covariates = none
h. Treatments = 3 treatments (top, middle, bottom)

2.16

a. Factors: Type of drug
b. Factor levels: D_1, D_2, Placebo
c. Blocks: Hospitals
d. Experimental units: Wards
e. Measurement units: Patients
f. Replications: 2 wards per drug in each of the 10 hospitals
g. Covariates: None
h. Treatments: D_1, D_2, Placebo

2.17

a. Factors: Type of treatment
b. Factor levels: D_1, D_2, Placebo
c. Blocks: Hospitals, Wards
d. Experimental units: Patients
e. Measurement units: Patients
f. Replications: 2 patients per drug in each of the ward/hospital combinations
g. Covariates: None
h. Treatments: D_1, D_2, Placebo

2.18

a. Factors: Type of school
b. Factor levels: Public; Private – non-parochial; Parochial
c. Blocks: Geographic region
d. Experimental units: Classrooms
e. Measurement units: Students in classrooms
f. Replications: 2 classrooms per each type of school in each of the city/region combinations
g. Covariate: Measure of socio-economic status
h. Treatments: Public; Private – non-parochial; Parochial

2.19

a. Factors: Temperature, Type of seafood
b. Factor levels: Temperature (0 °C, 5 °C, 10 °C); Type of seafood (oysters, mussels)
c. Blocks: None
d. Experimental units: Package of seafood
e. Measurement units: Sample from package
f. Replications: 3 packages per temperature
g. Treatments: (0 °C, oysters), (5 °C, oysters), (10 °C, oysters), (0 °C, mussels), (5 °C, mussels), (10 °C, mussels)

2.20

- Randomized complete block design with blocking variable (10 orange groves) and 48 treatments in a $3 \times 4 \times 4$ factorial structure.
- Experimental Units: Plots
- Measurement Units: Trees

2.21

- Randomized complete block design with blocking variable (10 warehouses) and 5 treatments (5 vendors)

2.22

- Randomized complete block design, where blocked by day
- 2-factor structure (where the factors are type of glaze, and thickness)

2.23

a. Design B. The experimental units are not homogeneous since one group of consumers gives uniformly low scores and another group gives uniformly high scores, no matter what recipe is used.

Using design A, it is possible to have a group of consumers that gives mostly low scores randomly assigned to a particular recipe. This would bias this particular recipe. Using design B, the experimental error would be reduced since each consumer would evaluate each recipe. That is, each consumer is a block and each of the treatments (recipes) is observed in each block. This results in having each recipe subjected to consumers who give low scores and to consumers who give high scores.

b. This would not be a problem for either design. In design A, each of the remaining 4 recipes would still be observed by 20 consumers. In design B, each consumer would still evaluate each of the 4 remaining recipes.

2.24

a. "Employee" should refer to anyone who is eligible for *sick days*.
b. Use payroll records. Stratify by employee categories (full-time, part-time, etc.), employment location (plant, city, etc.), or other relevant subgroup categories. Consider systematic selection within categories.
c. Sex (women more likely to be care givers), age (younger workers less likely to have elderly relatives), whether or not they care for elderly relatives now or anticipate doing so in the near future, how many hours of care they (would) provide (to define "substantial"), etc. The company might want to explore alternative work arrangements, such as flex-time, offering employees 4 ten-hour days, cutting back to 3/4-time to allow more time to care for relatives, etc., or other options that might be mutually beneficial and provide alternatives to taking sick days.

2.25

a. Each state agency and some federal agencies have records of licensed physicians, professional corporations, facility licenses, etc. Professional organizations such as the American Medical Association, American Hospital Administrators Association, etc., may have such lists, but they may not be as complete as licensing records.
b. What nursing specialties are available at this time at the physician's offices or medical facilities? What medical specialties/facilities do they anticipate adding or expanding? What staffing requirements are unfilled at this time or may become available when expansion occurs? What is the growth/expansion time frame?
c. Licensing boards may have this information. Many professional organizations have special categories for members who are unemployed, retired, working in fields not directly related to nursing, students who are continuing their education, etc.
d. Population growth estimates may be available from the Census Bureau, university economic growth research, bank research studies (prevailing and anticipated load patterns), etc. Health risk factors and location information would be available from state health departments, the EPA, epidemiological studies, etc.
e. Licensing information should be stratified by facility type, size, physician's specialty, etc., prior to sampling.

2.26

If phosphorous first: [P,N]

[10,40], [10,50], [10,60], then [20,60], [30,60] or
[20,40], [20,50], [20,60], then [10,60], [30,60] or
[30,40], [30,50], [30,60], then [10,60], [10,60]

If nitrogen first: [N,P]

[40,10], [40,20], [40,30], then [50,30], [60,30] or
[50,10], [50,20], [50,30], then [40,30], [60,30] or
[60,10], [60,20], [60,30], then [40,30], [50,30]

So, for example

Phosphorus	30	30	30	10	20
Nitrogen	40	50	60	60	60
Yield	150	170	190	165	185

Recommendation: Phosphorus at 30 pounds, and Nitrogen at 60 pounds.

2.27

	Factor 2		
Factor 1	I	II	III
A	25	45	65
B	10	30	50

2.28

a. Group dogs by sex and age:

Group	Dog
Young female	2, 7, 13, 14
Young male	3, 5, 6, 16
Old female	1, 9, 10, 11
Old male	4, 8, 12, 15

b. Generate a random permutation of the numbers 1 to 16:
15 7 4 11 3 13 8 1 12 16 2 5 6 10 9 14
Go through the list and the first two numbers that appear in each of the four groups receive treatment L_1 and the other two receive treatment L_2.

Group	Dog-Treatment
Young female	2-L_2, 7-L_1, 13, 14-L_2
Young male	3-L_1, 5-L_2, 6-L_1, 16-L_2
Old female	1-L_1, 9-L_2, 10-L_2, 11-L_1
Old male	4-L_1, 8-L_2, 12-L_2, 15-L_1

2.29

a. Bake one cake from each recipe in the oven at the same time. Repeat this procedure r times. The baking period is a block with the four treatments (recipes) appearing once in each block. The four recipes should be randomly assigned to the four positions, one cake per position. Repeat this procedure r times.
b. If position in the oven is important, then position in the oven is a second blocking factor along with the baking period. Thus, we have a Latin square design. To have $r = 4$, we would need to have each recipe appear in each position exactly once within each of four baking periods. For example:

Period 1		Period 2		Period 3		Period 4	
R_1	R_2	R_4	R_1	R_3	R_4	R_2	R_3
R_3	R_4	R_2	R_3	R_1	R_2	R_4	R_1

c. We now have an incompleteness in the blocking variable period since only four of the five recipes can be observed in each period. In order to achieve some level of balance in the design, we need to select enough periods in order that each recipe appears the same number of times in each period and the same total number of times in the complete experiment. For example, suppose we wanted to observe each recipe $r = 4$ times in the experiment. If would be necessary to have 5 periods in order to observe each recipe 4 times in each of the 4 positions with exactly 4 recipes observed in each of the 5 periods.

Period 1		Period 2		Period 3		Period 4		Period 5	
R_1	R_2	R_5	R_1	R_4	R_5	R_3	R_4	R_2	R_3
R_3	R_4	R_2	R_3	R_1	R_2	R_5	R_1	R_4	R_5

2.30

a. The 223 plots of approximately equal sized land from Google Earth (excluding water)
b. If there is some reason to believe the trees in the 'watery' regions differ from those in the other regions, this discrepancy may cause a divide in our sampling frame and the population of all trees in the region.
c. Again, if trees in the watery region tend to have larger trunk diameter, we would underestimate the number of trees with diameter of 12 inches or more.

2.31

a. All cars (and by extension, their tires) in the state.
b. Cars registered in the 4 months in which the sample was taken.
c. 2 potential concerns arise: not all cars in the region are registered and the time of year may lead to ignoring some cars (some people leave the area for the winter). Unregistered cars may have a higher proportion of unsafe tire tread thickness.

2.32

a. All corn fields in the state.
b. All corn fields in the state (if a list is available).
c. Stratified sampling plan in which the number of acres planted in corn determine the strata.
d. No biases appear present.

2.33

a. People are notoriously bad at recall. A telephone interview immediately following the time of interest would likely be best, but nonresponse is often high. Mailed questionnaires would likely be administered too late to be of use and personal interviewing would be intractable to interview in a timely manner.
b. All three are potential avenues. Interviews are more personal but more time consuming. Mailing questionnaires should also work as the editor has a list of his/her clientele, but if he wants to garner information about perspectives of those not reading his/her paper, he/she may need to blanket the city with questionnaires. Telephone interviews may be difficult as finding the numbers of those in the area may be difficult.
c. Again, all three methods would be viable. A mailed questionnaire would be the easiest and cheapest but the response rate would likely be lower.
d. If the county believes they have an accurate list of those with dogs, a mailed questionnaire or telephone interview would work, but using a list of registered dogs may be underrepresenting those who haven't taken good care of their dogs (and thereby underrepresenting the proportion with rabies shots).

2.34

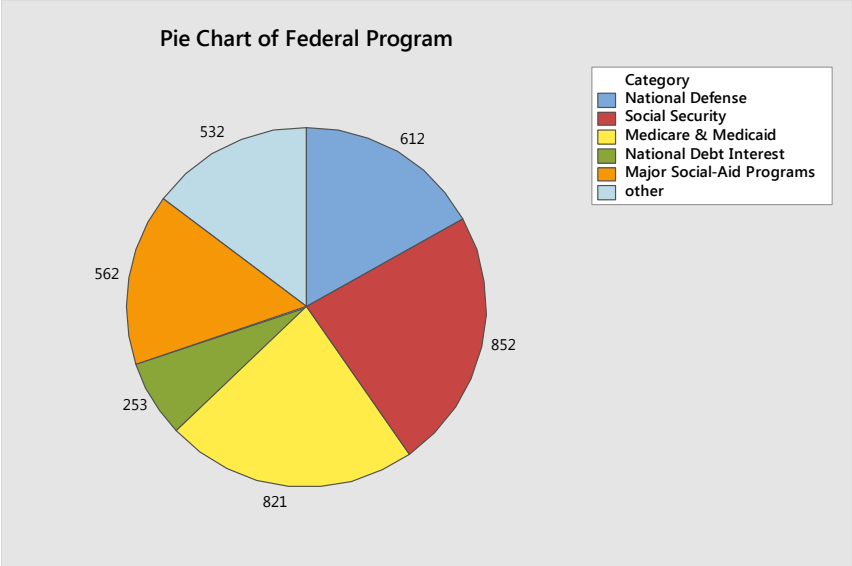
- People who cheat on their taxes are unlikely to admit to it readily. Therefore, the poll likely underestimates the true percentage of people who cheat on their taxes. Garnering truthful responses, even if anonymity is guaranteed, on questions of a personal nature can be a challenge.

Chapter 3

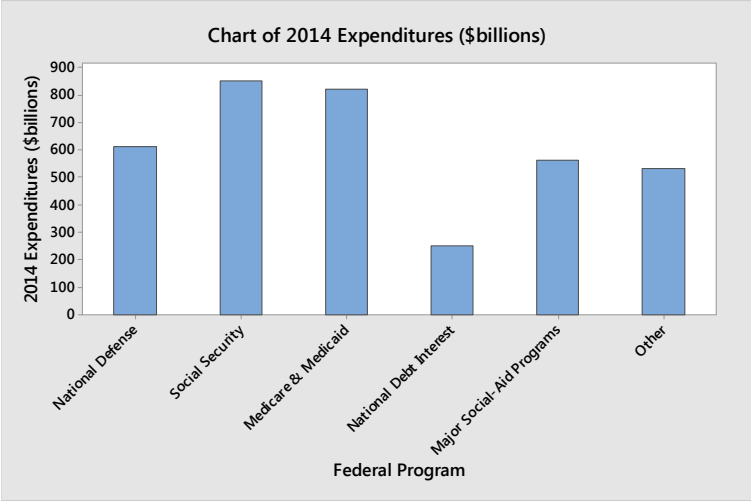
Data Description

3.1

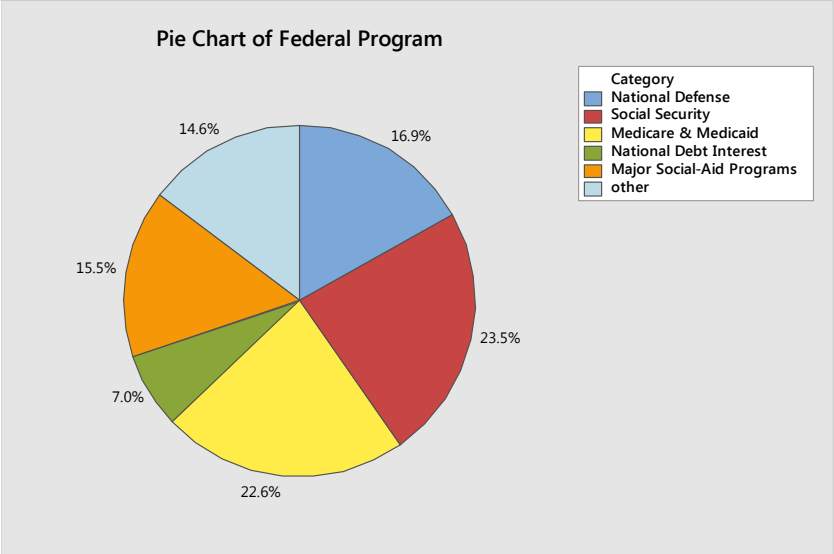
a. The following is a pie chart of the federal expenditures for the 2014 fiscal year (in billions of dollars).

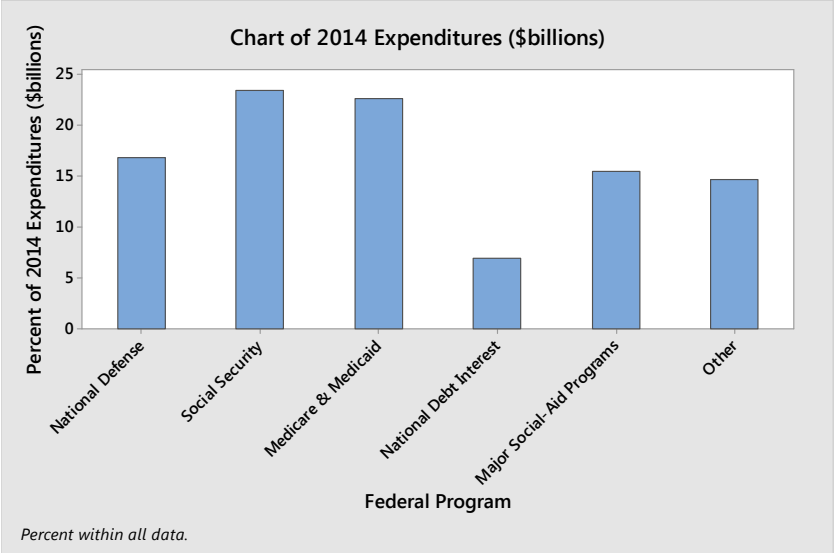


b. The following is a bar chart of the federal expenditures for the 2006 fiscal year (in billions of dollars).



c. The following are a pie chart and bar chart of the federal expenditures for the 2014 fiscal year (in percentages).



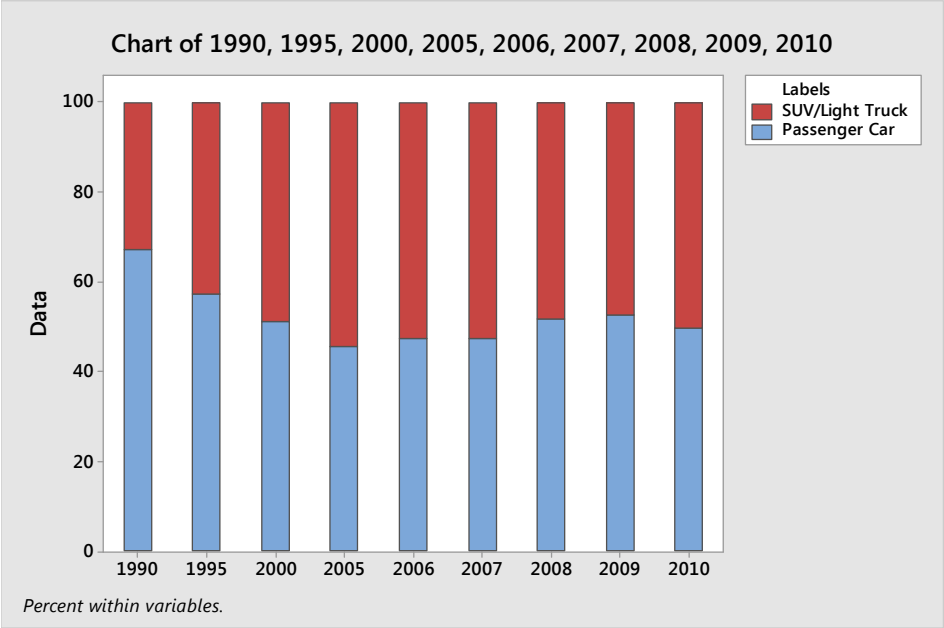


d. The pie chart using percentages is probably most informative to the tax-paying public. Here the tax-paying public can compare the percentages spent by the Federal government for domestic and defense programs as part of a whole.

3.2

a. Pie charts would not be appropriate to display these data. We would not be able to see trends over time.

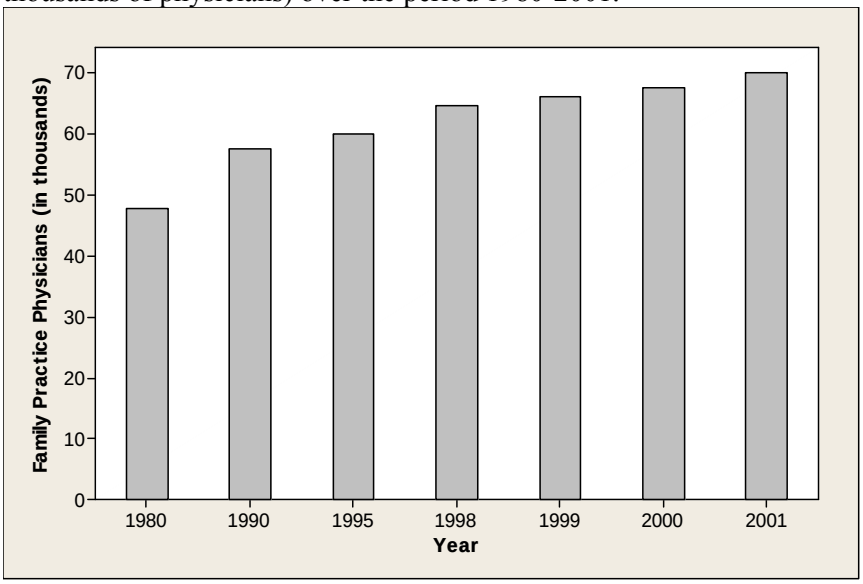
b. The following bar chart shows the changes across the 20 years in the public's choice in vehicle.



c. It appears that the percentage of passenger cars has decreased over the period 1990-2010. If there was a substantial increase in gasoline prices, we would expect the percentage of passenger cars to increase.

3.3

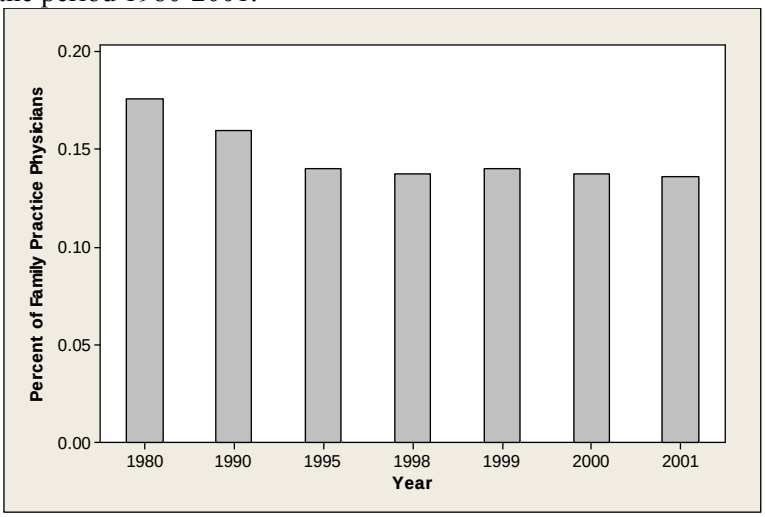
a. The following bar chart shows the increase in the number of family practice physicians (in thousands of physicians) over the period 1980-2001.



b. The percent of office-based physicians who are family practice physicians over the period 1980-2001 can be seen in the following table.

	1980	1990	1995	1998	1999	2000	2001
Percent Family Practice	17.6	16.0	14.0	13.8	14.0	13.8	13.6

The following bar chart shows the percent of office-based physicians who are family practice physicians over the period 1980-2001.

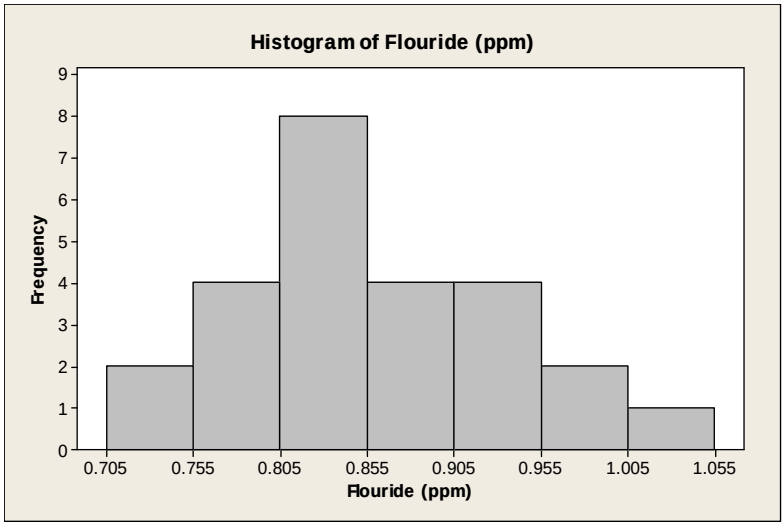


c. While the number of family practice physicians increased over the period 1980-2001, the percent of total office-based physicians who are family practice physicians decreased over the same period.

3.4

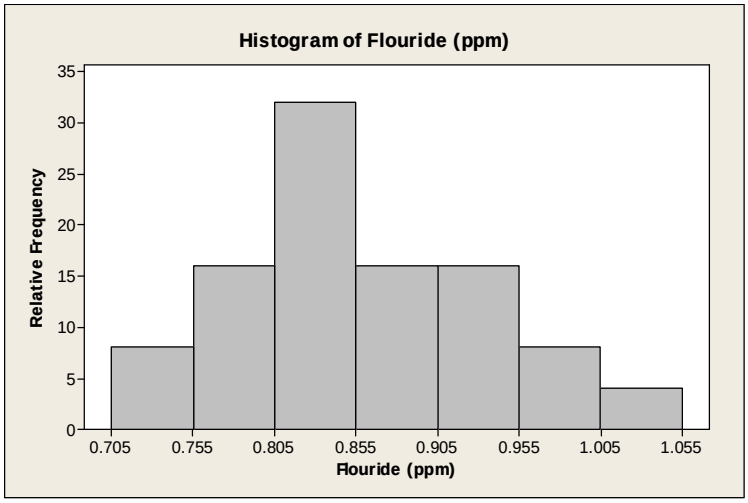
a. Range = 1.05 – 0.72 = 0.33

b. The frequency histogram should be plotted with 7 classes ranging from 0.705 to 1.055. The intervals should have width 0.05.



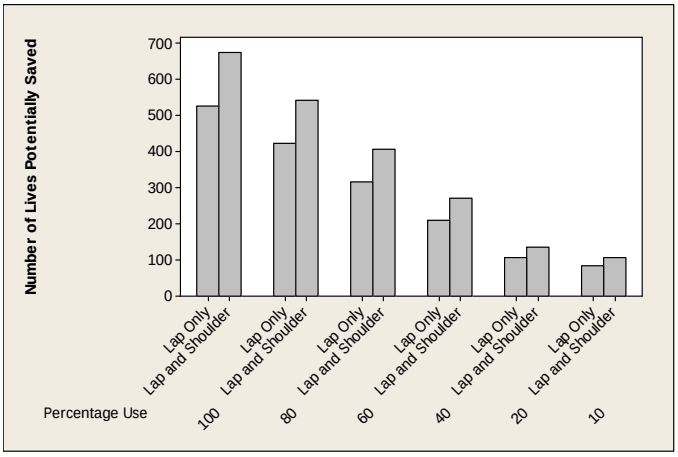
c. Relative frequencies are given below. Plot relative frequencies versus class intervals.

Class	Class Interval	Frequency (f_i)	Relative Frequency (f_i/25)
1	0.705-0.755	2	0.08
2	0.755-0.805	4	0.16
3	0.805-0.855	8	0.32
4	0.855-0.905	4	0.16
5	0.905-0.955	4	0.16
6	0.955-1.005	2	0.08
7	1.005-1.055	1	0.04
Total		$n = 25$	1.00



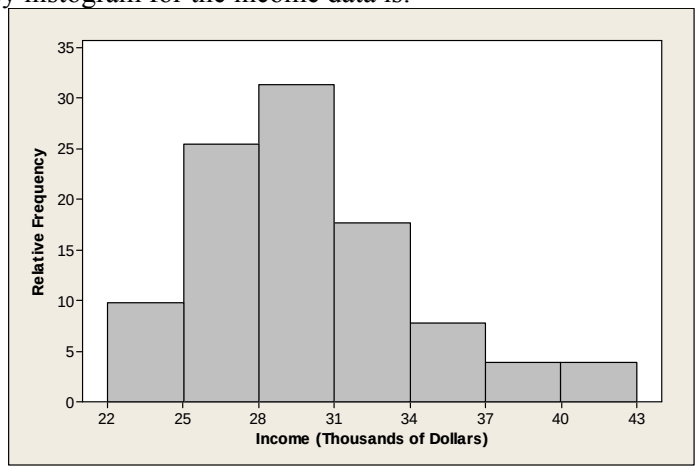
d. The probability is 7/25 = 0.28 that the fluoride reading would be greater than 0.90 ppm. Thus, we would predict that 28% of the days would have a reading greater than 0.90 ppm.

3.5 Two separate bar graphs could be plotted, one with Lap Belt Only and the other with Lap and Shoulder Belt. A single bar graph with the Lap Belt Only value plotted next to the Lap and Shoulder Belt for each value of Percentage of Use is probably the most effective plot. This plot would clearly demonstrate that the increase in the number of lives saved by using a shoulder belt increased considerably as the percentage use increased.



3.6

a. A relative frequency histogram for the income data is:

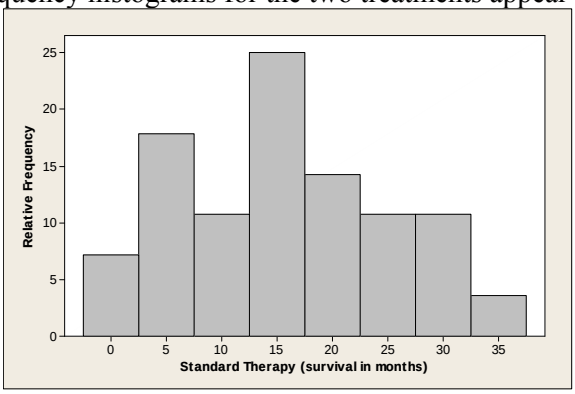


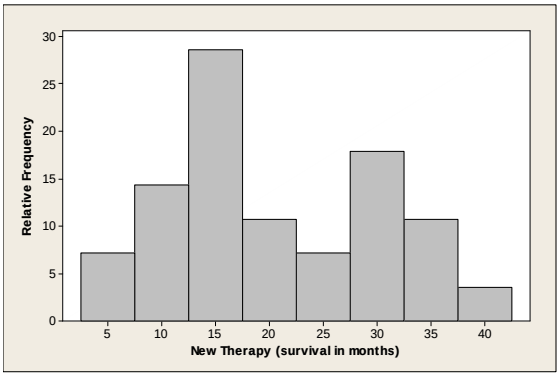
b. The histogram is unimodal and skewed to the right. The median is in the $31,000-$33,900 bin. There are no outliers.

c. Since there is a large spread in incomes (about $21,000), the data are not homogeneous across the states.

3.7

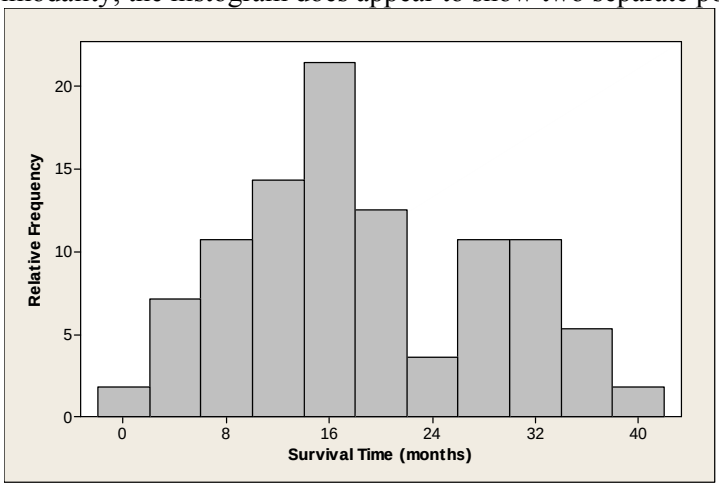
a. The separate relative frequency histograms for the two treatments appear below.





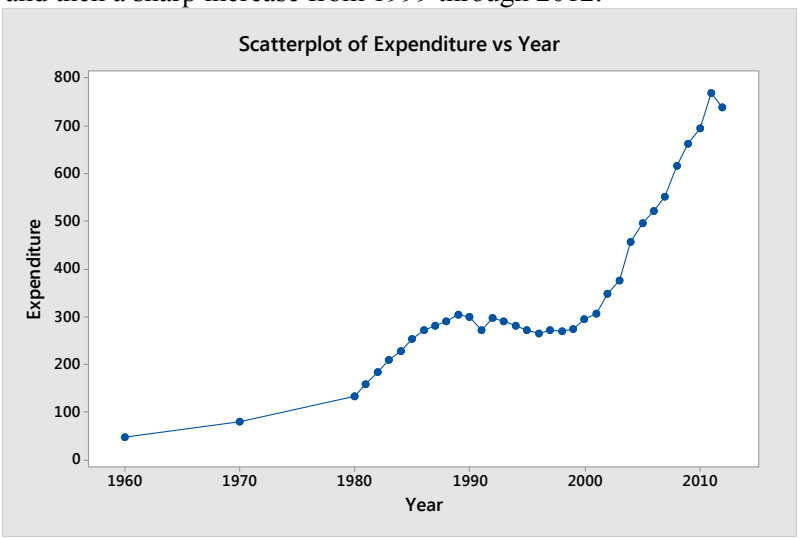
b. The histogram for the New Therapy begins and ends with bins that are slightly higher than the bins in the histogram for the Standard Therapy. This would indicate that the New Therapy generates a few more large values than the Standard Therapy. However, there is not convincing evidence that the New Therapy generates a longer survival time.

3.8 The following histogram of the data from the two therapies combined is bimodal and skewed to the right. Because of the bimodality, the histogram does appear to show two separate populations



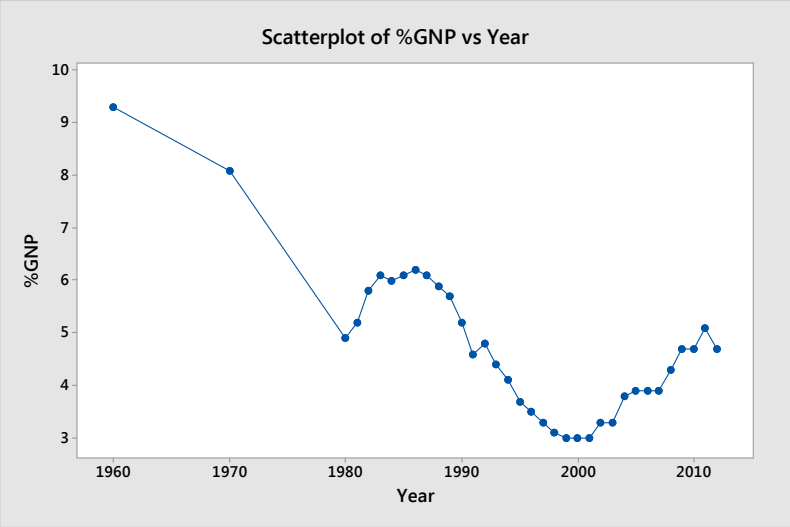
3.9

a. The time series plot shows an increase from 1980 through 1990, with a large dip at 1991 (after the Persian Gulf Conflict). There is then a decrease in expenditures in billions of dollars over the period 1992 to 1999, and then a sharp increase from 1999 through 2012.



b. The time series plot of expenditures as a percent of GNP shows a large decrease from 1960-1980 and an increase from 1980 to 1983. It is then fairly steady from 1983 to 1987, decreases from 1987

to 2001 (with the exception of a spike in 1992), has a sharp increase 2001 to 2002, and then is fairly steadily increasing from 2002 to 2012.

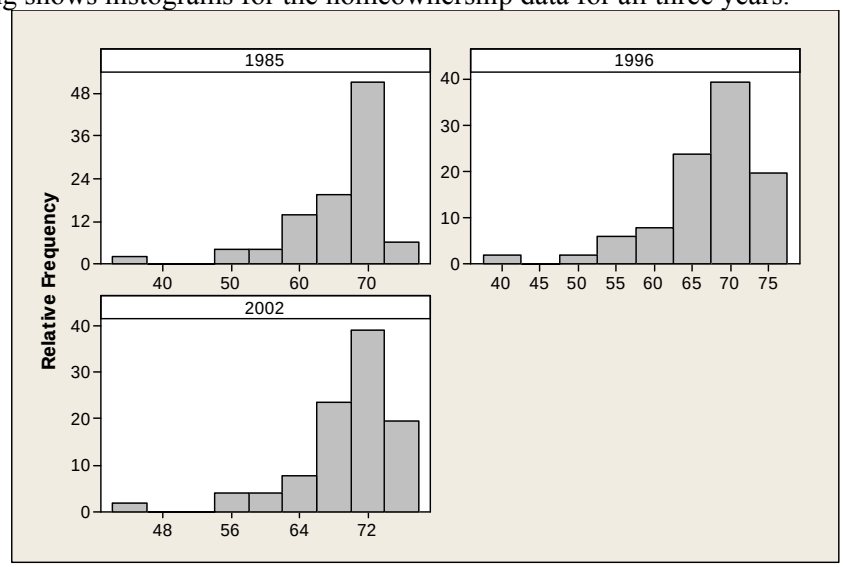


c. The plots do not show similar trends. The time series plot of expenditures supports the assertions of these members of Congress, but the %GDP vs. Year plot appears to contradict those liberal members.

d. The expenditures show a steady rise due to inflation. The %GDP plot is much more telling of the landscape. % military spending show increases that can be related to various historical events. The increase in the early 80s is likely due to the Iran Contra affair. 91-92 shows a spike for Operation Desert Storm. The recent increasing trend is likely explained by the War on Terror.

3.10

a. The following shows histograms for the homeownership data for all three years.



b. All three histograms are unimodal and skewed to the left. It appears from the histograms that homeownership has increased over the years.
c. Perhaps more people are able to afford to live in their own homes instead of renting homes from others.
d. Congress could address the issues that a higher proportion of homes are owner-occupied. Congress could then develop tax laws for owner-occupied homes.

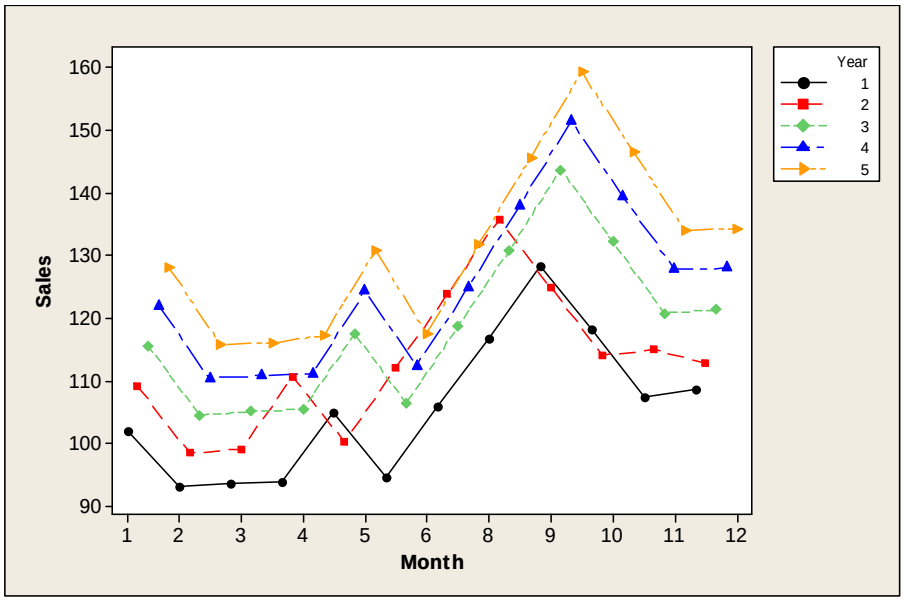
3.11 The following are stem-and-leaf plots for each of the three years 1985, 1996, and 2002.

	1985		**1996**		**2002**
3	7	3		3	
4		4	0	4	4
4		4		4	
5	014	5	02	5	
5	7	5	56	5	5789
6	00011122334	6	1112233444	6	23
6	5566677777888888999999	6	5666777788888888999	6	55677778899999
7	000000111233	7	000111122233344	7	000000011122222333334444
7	5	7	56	7	556777

3|7 = 37% (stems are tens, leaves are ones)

3.12 All three homeownership distributions are unimodal and skewed to the left (since they are skewed, they are not symmetric). The homeownership distributions each have outliers (37% in 1985, 40% in 1996, and 44% in 2002).

3.13 The plots show an upward trend from year 1 to year 5. There is a strong seasonal (cyclic) effect; the number of units sold increases dramatically in the late summer and fall months.



3.14 The mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^{20} y_i = \frac{155+25+30+\cdots+69}{20} = 60.6$. The median is the average of the 10th and 11th values when arranged in increasing order: median = 32.5. The mode is 29.

3.15 The mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^{21} y_i = \frac{35+81+96+\cdots+24}{21} = 55.19$. The median is the 11th value when arranged in increasing order: median = 58. The modes are 24 and 58.

3.16 The mean changes to 86.6 and the median and mode are unchanged. Median and mode are robust to outliers

3.17

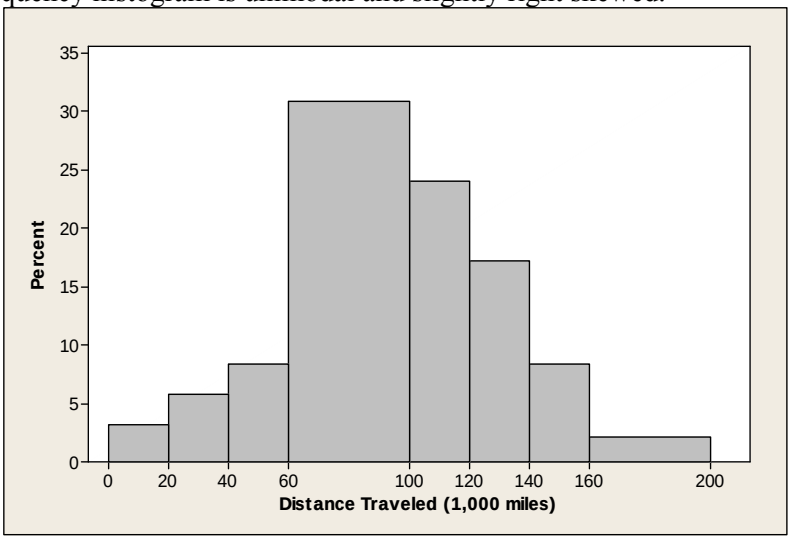
	10% Trimmed Mean	5% Trimmed Mean
3.15	53.5647	54.9789
3.16	53.5647	68.2421

The 10% trimmed mean for 3.15 and 3.16 are the same because the outliers are both trimmed off. The 5% trimmed mean differs as only the highest and lowest value are trimmed so one of the outliers is still accounted for in the 5% trimmed mean.

3.18 The mean is 9.5, the median is 9.0556 using the group data median formula, and the mode are is 9.5.

3.19

a. The relative frequency histogram is unimodal and slightly right skewed.



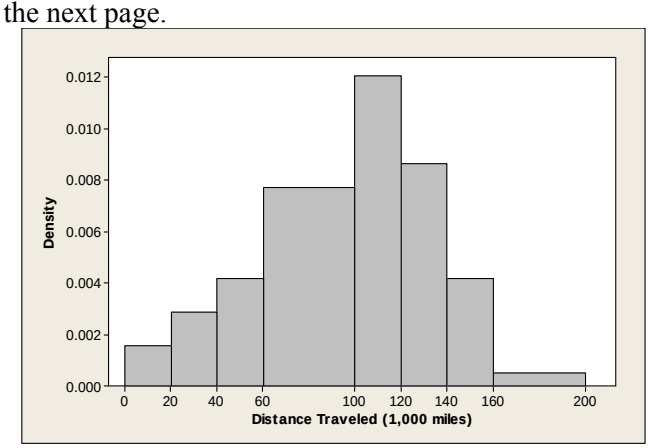
b. The following table is used to calculate the summary statistics:

Class Interval	Frequency (f_i)	Midpoint (y_i)	$f_i y_i$
0-20.0	6	10	60
20.1-40.0	11	30	330
40.1-60.0	16	50	800
60.1-100.0	59	80	4720
100.1-120.0	46	110	5060
120.1-140.0	33	130	4290
140.1-160.0	16	150	2400
160.1-200.0	4	180	720
Total	191		18,380

mean ≈ 18380/191 = 96.2
mode ≈ 80

$$\text{median} \approx L + \frac{w}{f_m}(0.5n - cf_b) = 100.1 + \frac{20}{46}\left[(0.5)(191) - 92\right] = 101.6$$

c. Since the median is larger than the mean, it would indicate that the plot is somewhat left skewed. This contradiction between what is indicated in the relative frequency histogram and what is indicated by the summary statistics is due to the fact that the class intervals are a different width. The correct plot would have relative frequency divided by class width on the vertical axis. This would then produce a left skewed histogram with mode at approximately 110. That histogram appears at the top of the next page.



d. The median is more informative since the distribution is somewhat skewed to the left which produces a mean somewhat less than the middle of the distribution. The median distance traveled would at least represent a value such that half of the buses traveled less and half greater than 101,600 miles.

3.20

a. The mean cannot be approximated since we do not know the endpoint for the last interval, hence cannot compute the midpoint for this interval.
Mode ≈ 240.

$$\text{median} \approx L + \frac{w}{f_m}(0.5n - cf_b) = 200 + \frac{19.9}{88}\left[(0.5)(408) - 119\right] = 219.2$$

b. The median since it would give an indication of the cholesterol level for which half of the men in this group have a greater or lesser value.

3.21

a. Mean = 8.04, Median = 1.54
b. Terrestrial: Mean = 15.01, Median = 6.03
   Aquatic: Mean = 0.38, Median = 0.375
c. The mean is more sensitive to extreme values than is the median.
d. Terrestrial: Median, because the two large values (76.50 and 41.70) result in a mean that is larger than 82% of the values in the data set.
   Aquatic: Mean or median since the data set is relatively symmetric.

3.22

a. After removing the survival times of the two individuals who left the study, we obtain Mean = 35.22 days. The median can be calculated for all 11 patients, since we know that the values for the two individuals who left the study were greater than the list values 57 and 60 which would place them in the upper half of the survival times. Thus, the Median = 29 days.
b. The median would be unchanged but the mean would increase since these two values will be greater than the mean calculated from the nine observed values.

3.23

a. If we use all 14 failure times, we obtain Mean > 173.7 days and Median = 154 days. In fact, we know that the mean is greater than 173.7 days since the failure times for two of the engines are greater than the reported times of 300 days.
b. The median would be unchanged if we replace the failure times of 300 with the true failure times for the two engines that did not fail. However, the mean would be increased.

3.24

a. The values are given below:

Group	Mean	Median	Mode
I	2.923	2.805	no mode
II	1.592	1.565	1.55, 1.57
III	0.797	0.755	0.70

b. Mean = 1.7707; Median = 1.565; Modes = 0.70, 1.55, 1.57
c. If we were to use one summary for the combined group, then the median would be most appropriate because the three groups are substantially different. If separate summaries are computer for each group, then the mean and median are both appropriate since the three groups have relatively symmetric distributions.

3.25 Mean = 1.7707, Median = 1.7083, Mode = 1.273
The average of the three net group means and the mean of the complete set of measurements are the same. This will be true whenever the groups have the same number of measurements, but it is not true if the groups have different sample sizes. However, the average of the group medians and modes are different from the overall median and mode.

3.26

a. $\bar{y} = \frac{\sum y_i}{n} = \frac{40}{8} = 5$ years. This value does appear to adequately represent the data set.

b.

$$\sum_{i=1}^{8} (y_i - 5)^2 = (6-5)^2 + (3-5)^2 + (10-5)^2 + (4-5)^2 + (4-5)^2 + (2-5)^2 + (4-5)^2 + (7-5)^2$$
$$= 1 + 4 + 25 + 1 + 1 + 9 + 1 + 4 = 46$$

c. $s^2 = \frac{46}{8-1} = 6.57 \Rightarrow s = 2.56$ years. The $CV = 100\frac{s}{y}\% = 100\frac{2.56}{5}\% = 51\%$. The standard deviation is 51% of the mean.

3.27

a. $s = 7.95$ years

b. Because the magnitude of the racers' ages is larger than that of their experience.

3.28

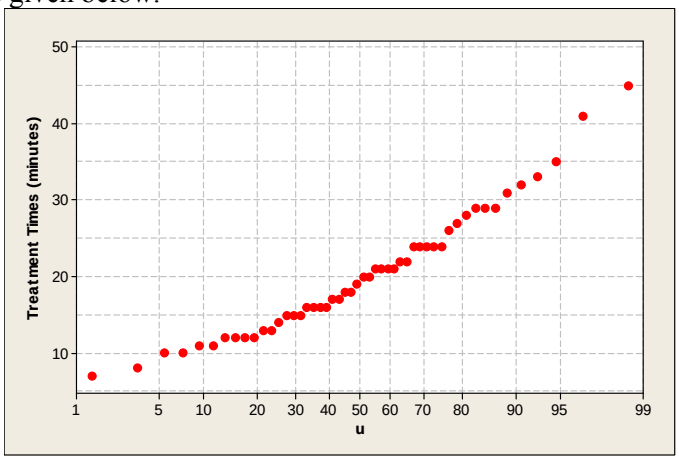
a. Racers' age: $CV = 100\frac{s}{y}\% = 100\frac{7.95}{29.875}\% = 26.6\%$

Years of experience: $CV = 100\frac{s}{y}\% = 100\frac{2.56}{5}\% = 51\%$

The CV for the racers' age is about half of the CV for their years of experience.

b. Estimated SD for racers' age: (39 – 18)/4 = 5.25
Estimated SD for years of experience: (10 – 2)/4 = 2
The estimate standard deviation for the years of experience is off by about 0.5 years, while the estimated standard deviation for the racers' ages is off by about 2.7 years.

3.29 The quantile plot is given below.



a. The 25th percentile is the value associated with $u = 0.25$ on the graph, which is 14 minutes. Also, by definition 14 minutes is the 25th percentile since 25% of the times are less than or equal to 14 minute and 75% of the times are greater than or equal to 14 minutes.

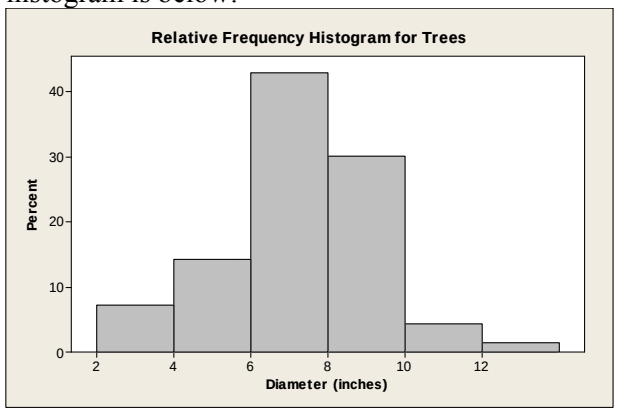
b. Yes; the 90th percentile is 31.5 minutes. This means that 90% of the patients have a treatment time less than or equal to 31.5 minutes (which is less than 40 minutes).

3.30

a. The frequency table is given here.

Class Intervals	Frequency	Relative Frequency
2.0-4.0	5	0.0714
4.1-6.0	10	0.1429
6.1-8.0	30	0.4286
8.1-10	21	0.3000
10.1-12.0	3	0.0429
12.1-14.0	1	0.0142
Total	70	1.0000

The relative frequency histogram is below:



b. $\bar{y} = 541/70 = 7.7286$

c. $s = 1.985$

$\bar{y} \pm s$ yields (5.744, 9.713); 50/70 = 71.43%

$\bar{y} \pm 2s$ yields (3.759, 11.698); 68/70 = 95.71%

$\bar{y} \pm 3s$ yields (1.774, 13.683); 70/70 = 100%

The percentages are very close to the percentages given by the Empirical Rule: 68%, 95%, and 99.7%.

3.31

a. Luxury: $\bar{y} = 145.0$, $s = 27.6$

Budget: $\bar{y} = 46.1$, $s = 5.13$

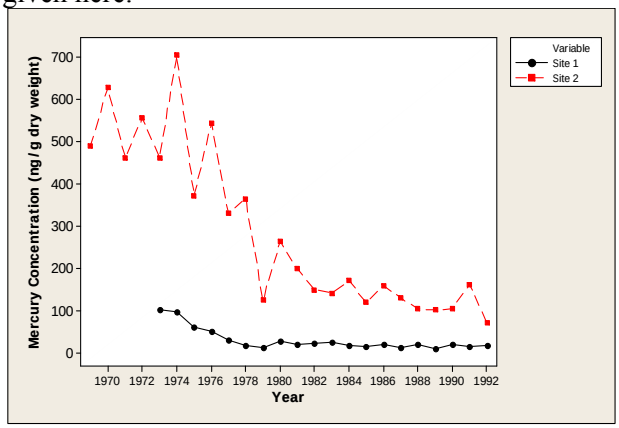
b. Luxury: CV = 19%
Budget: CV = 11%

c. Luxury hotels vary in quality, location, and price, whereas budget hotels are more competitive for the low-end market so prices tend to be similar.

d. The CV would be better because it takes into account the larger difference in the means between the two types of hotels.

3.32

a. The time series plot is given here.



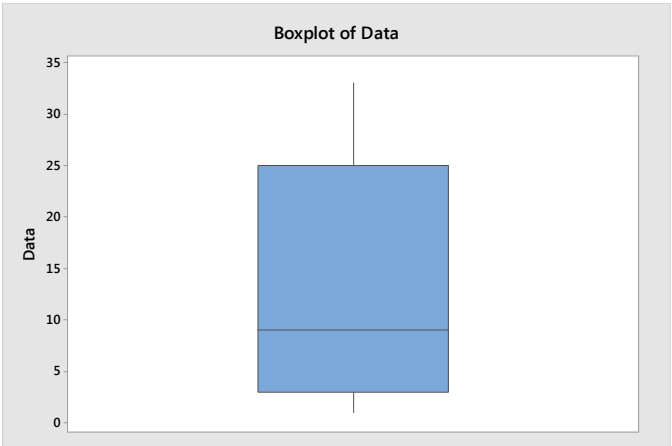
For Site 1, there is a steady decrease until 1980, after which the level is fairly constant but at a much lower level than the values for Site 2. The concentrations at Site 2 are very erratic from 1969 to 1980, with alternating rises and falls. From 1980 through 1992, there is a fairly steady decline in mercury concentration.

b. Site 1: Median = 18.25 ng/g; Mean = 29.18 ng/g
Site 2: Median = 184.1 ng/g; Mean = 287.1 ng/g
Both distributions are right skewed, thus the median is a more appropriate measure of center than is the mean. Site 1 has a considerably lower center than that of Site 2.

c. Site 1: s = 26.95 ng/g; CV = 92%
Site 2: s = 194.7 ng/g; CV = 68%
Comparing the standard deviations can be misleading because 26.95 is smaller in magnitude than 194.7. However, the data values for Site 1 are considerably smaller in magnitude as well. Therefore, it is more informative to compare the CV values. Based on CV values, the concentrations from Site 1 are relatively more variable than those from Site 2.

d. No, Site 1 does not have values for these years.

3.33

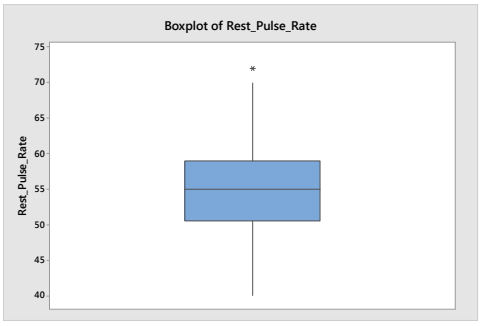


3.34

a. A stem-and-leaf plot is below:

Stem	Leaf
4	01
4	69999
5	1244
5	55555556677899
6	013
6	5
7	02

b. The boxplot is shown below.



c. The distribution appears fairly symmetric and unimodal.

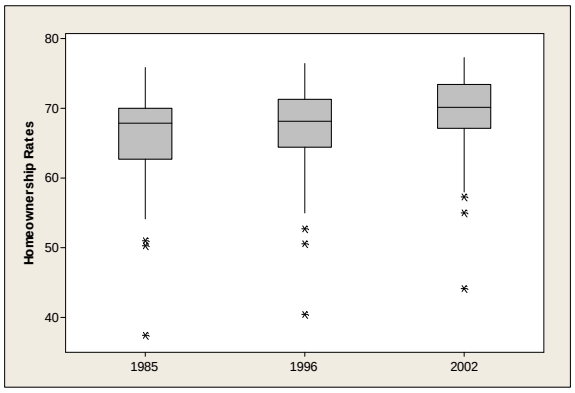
d. The population of interest is participants in a 10K race. Likely, these are somewhat 'fit' individuals.

3.35

a. CAN: $Q_1 \approx 1.45$, Q_2 = Median ≈ 1.65, $Q_3 \approx 2.4$
DRY: $Q_1 \approx 0.55$, Q_2 = Median ≈ 0.60, $Q_3 \approx 0.70$

b. Canned dog food is more expensive (median much greater than that for dry dog food), highly skewed to the right with a few large outliers. Dry dog food is slightly left skewed with a considerably less degree of variability than canned dog food.

3.36 The following shows comparative boxplots for homeownership rates for the years 1985, 1996, and 2002:



a. The distributions for each of the three years (1985, 1996, and 2002) are each left skewed with a few low outliers. The medians appear to be greater, and the distributions as a whole tend to be higher, in subsequent years.
b. The descriptions in part (a) agree with our description of the distributions in Exercise 3.11.

3.37

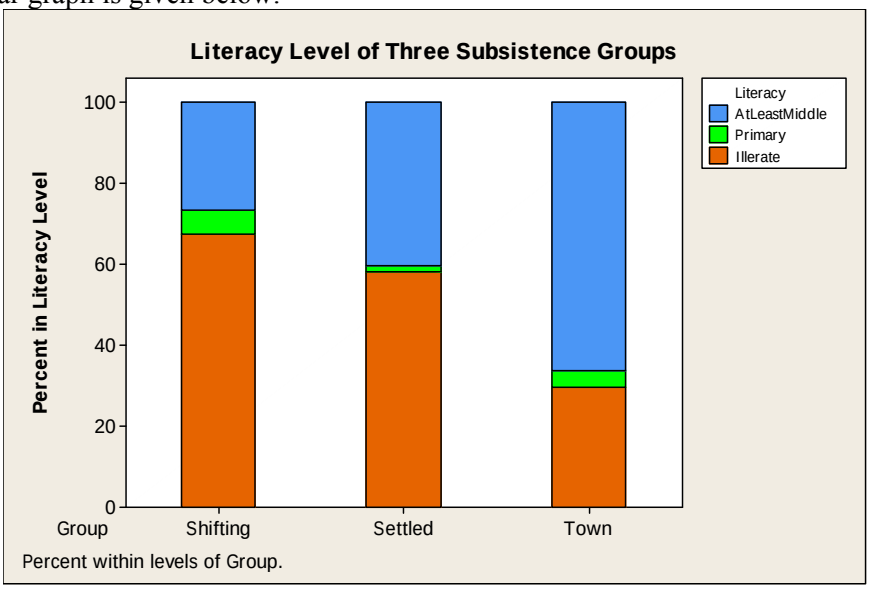
a. 1985: mean = 65.876; median = 67.90
1996: mean = 66.843; median = 68.20
2002: mean = 69.449; median = 70.20
Since the distributions are left skewed, it is better to use the median for each of the three years.
b. 1985: s = 6.734; CV = 10.2%
1996: s = 6.688; CV = 10.0%
2002: s = 6.163; CV = 8.9%
The coefficients of variation are decreasing over the three years.

3.38

a. See the boxplot in Exercise 3.36. Median homeownership rate is increasing over the three years.
b. See the boxplot in Exercise 3.36. The variation in homeownership rate is decreasing over the three years.
c. The District of Columbia, Hawaii, and New York have extremely low homeownership rates in each of the three years. These states are indicated as outliers in the side-by-side boxplot.
d. No states have extremely high homeownership rates in each of the three years.

3.39

a. A stacked bar graph is given below.



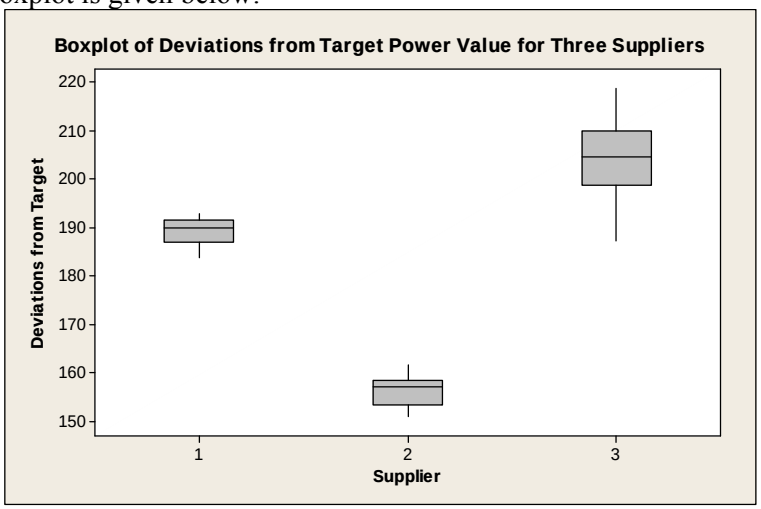
b. Illiterate: 46%; Primary Schooling: 4%; At Least Middle School: 50%
Shifting Cultivators: 27%; Settled Agriculturists: 21%; Town Dwellers: 51%
There is a marked difference in the distribution of the three literacy levels for the three subsistence groups. Town dwellers and shifting cultivators have the reverse trends in the three categories, whereas settled agriculturists fall into essentially two classes.

3.40

a. The means and standard deviations are given below:

Supplier	$\bar{y}$	s
1	189.23	2.96
2	156.28	3.30
3	203.94	8.98

b. A side-by-side boxplot is given below:

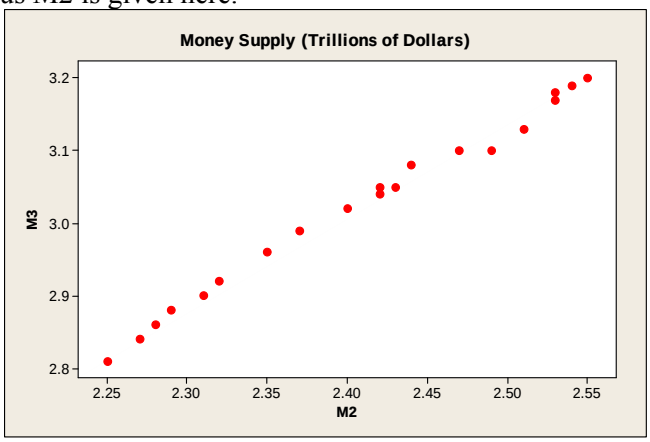


The three distributions are relatively symmetric, but supplier 3 is considerably more variable and is shifted about supplier 1's values, which in turn are shifted above supplier 2's values.

c. Supplier 3 not only has the largest mean but also the largest standard deviation. Suppliers 1 and 2 have similar degrees of variability, but supplier 1 has a greater mean than supplier 2.

d. Supplier 2 because it has the smallest mean and deviated with essentially the same degree of variability as supplier 1.

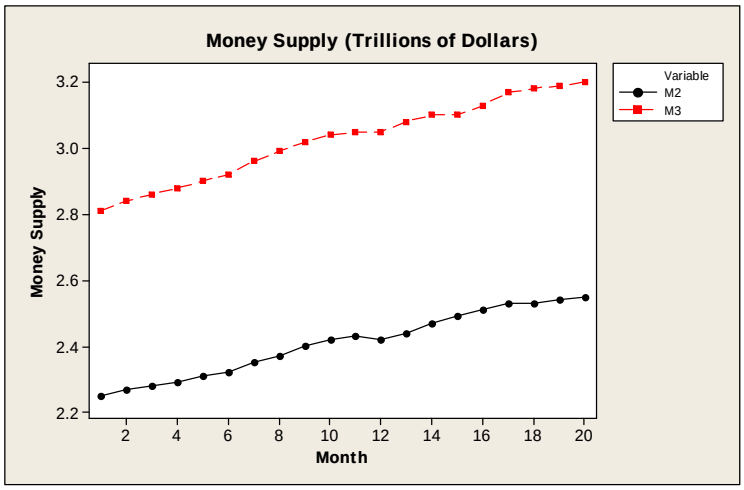
3.41 A scatterplot of M3 versus M2 is given here.



a. Yes, it would since we want to determine the relative changes in the two over the 20-month period of time.

b. See scatterplot. The two measures follow an increasing, approximately linear relationship.

3.42 A time series plot with M2 and M3 values on the vertical axis and months on the horizontal axis is given here.



This is a more informative plot than the scatterplot because it shows the relative changes of the two measures of money supply across the 20 months.

3.43

a. Mean = 57.5; Median = 34.0
b. Median since the data has a few very large values which results in the mean being larger than all but a few of the data values.
c. Range = 273; $s = 70.2$
d. Using the approximation, $s \approx$ range/4 = 273/4 = 68.3. The approximation is fairly accurate.
e. $\bar{y} \pm s \Rightarrow (-12.7, 127.7)$; yields 82%
   $\bar{y} \pm 2s \Rightarrow (-82.9, 197.0)$; yields 94%
   $\bar{y} \pm 3s \Rightarrow (-153.1, 268.1)$; yields 97%
f. The Empirical Rule applies to data sets with roughly a "mound-shaped" histogram. The distribution of this data set is highly skewed right.

3.44

a. Mode = 5 hours; Median = 15 hours; Mean = 15.96 hours
b. Range = 34 – 4 = 30 hours; $s \approx 30/4 = 7.5$ hours
c. $s = 8.5$ hours
d. No, the histogram for the data set is skewed to the right and hence is not mound-shaped.

3.45

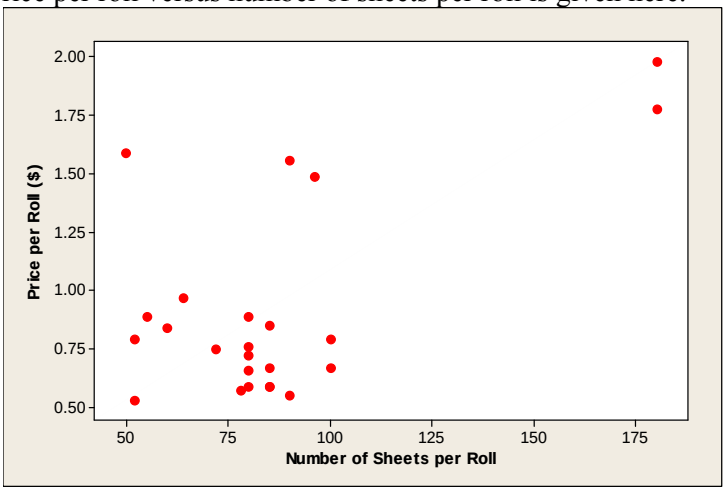
a. Price per roll: Mean = 0.9196, $s = 0.4233$
   Price per sheet: Mean = 0.01091, $s = 0.0059$
b. Price per roll: $\text{CV} = 100\dfrac{0.4233}{0.9196}\% = 46.03\%$

Price per sheet: $CV = 100\dfrac{0.0059}{0.01091}\% = 54.13\%$

The price per sheet is more variable relative to its mean.

c. CV; The CV is unit free, whereas the standard deviation also reflects the relative magnitude of the data values.

3.46 A scatterplot of price per roll versus number of sheets per roll is given here.

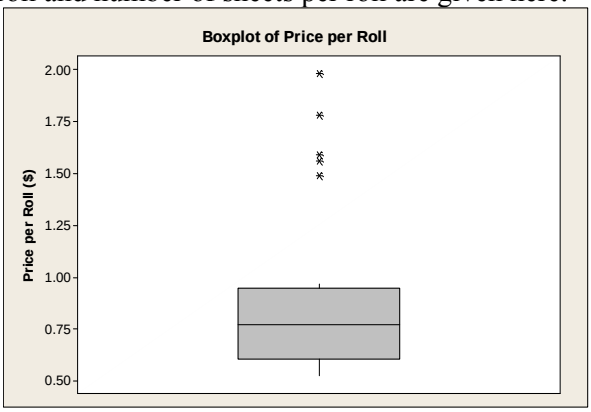


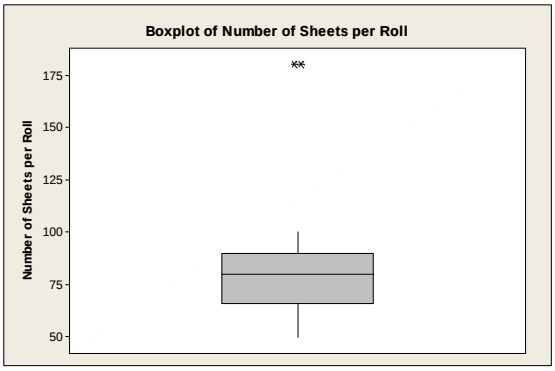
a. No.

b. No, as the number of sheets increases from 50 to 100, there is just a scatter of points, no real pattern. The price per roll jumps dramatically for the two brands having the largest number of sheets.

c. Paper towel sheets vary in thickness and size, both of which will affect the price.

3.47 Boxplots for price per roll and number of sheets per roll are given here.

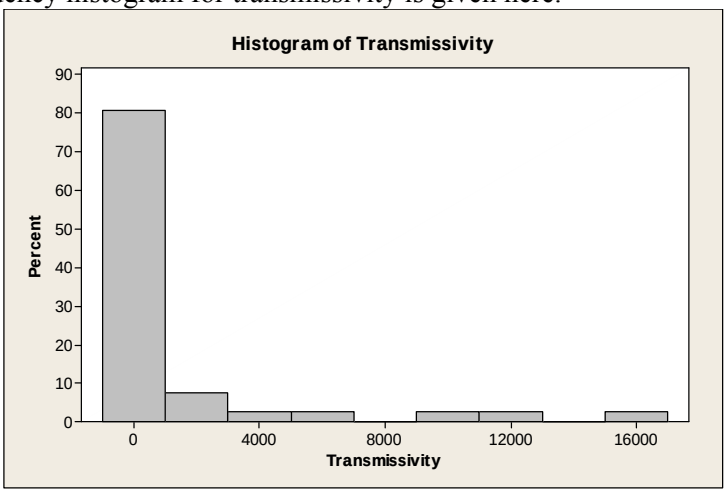




From the two boxplots, there a 5 unusual brands with regard to price per roll: \$1.49, \$1.56, \$1.59, \$1.78, and \$1.98. There are 2 unusual brands with respect to number of sheets per roll: 180 and 180.

3.48

a. A relative frequency histogram for transmissivity is given here.



b. The distribution is unimodal and highly skewed to the right.

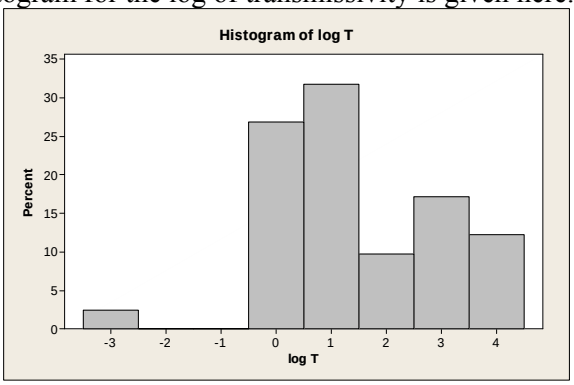
c. $\bar{y} \pm s = 1424 \pm 3488 \Rightarrow (-2063, 4912)$ contains 37/41 = 90.2%

$\bar{y} \pm 2s = 1424 \pm (2)3488 \Rightarrow (-5551, 8400)$ contains 38/41 = 92.7%

$\bar{y} \pm 3s = 1424 \pm (3)3488 \Rightarrow (-9039, 11888)$ contains 39/41 = 95.1%

These values do not match the values from the Empirical Rule: 68%, 95%, and 99.7%.

d. A relative frequency histogram for the log of transmissivity is given here.



The shape is more mound-shaped than the original data, although it appears somewhat skewed to the right with a low outlier.

$\bar{y} \pm s = 1.48 \pm 1.54 \Rightarrow (-0.06, 3.02)$ contains 31/41 = 75.6%

$\bar{y} \pm 2s = 1.48 \pm (2)1.54 \Rightarrow (-1.60, 4.56)$ contains 40/41 = 97.6%

$\bar{y} \pm 3s = 1.48 \pm (3)1.54 \Rightarrow (-3.14, 6.10)$ contains 41/41 = 100%

These values more closely match the percentages from the Empirical Rule.

3.49 A relative frequency histogram for murder rate is given here.

3.50

a. Mode = 2.5; Median $\approx L + \frac{w}{f_m}\left[0.5n - cf_b\right] = 5.5 + \frac{2}{13}\left[0.5(90) - 35\right] = 7.04$

Mean $\approx \frac{1}{n}\sum_{i=1}^{13} f_i y_i = 747/90 = 8.3$

b. Since the distribution is skewed to the right, the median provides a better measure of the center of the distribution.

3.51

a. 75th percentile $\approx L+\frac{w}{f_m}\left[0.75n-cf_b\right]=13.5+\frac{2}{9}\left[0.75(90)-72\right]=12.5$

25th percentile $\approx L+\frac{w}{f_m}\left[0.25n-cf_b\right]=3.5+\frac{2}{15}\left[0.25(90)-20\right]=3.83$

IQR $\approx 12.5 - 3.83 = 8.67$

$$s^2 \approx \frac{1}{n-1}\sum_{i=1}^{13} f_i(y_i-8.3)^2$$

$$=\frac{1}{89}\left[2(0.5-8.3)^2+18(2.5-8.3)^2+\cdots+1(24.5-8.3)^2\right]=29.0382$$

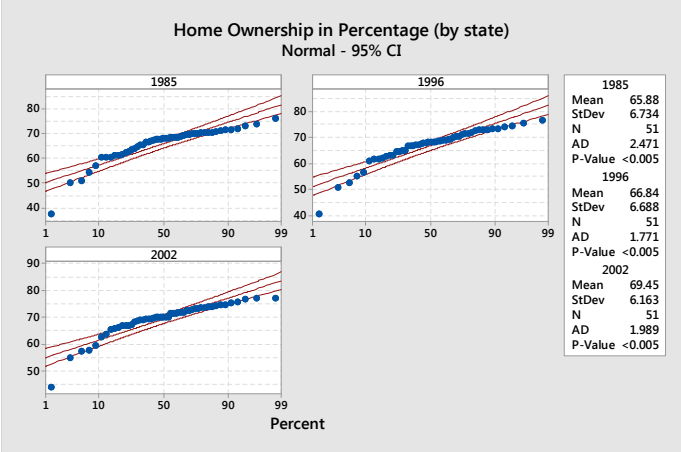
Thus, $s \approx \sqrt{29.0382} = 5.389$.

b. Because the distribution is skewed right, the IQR is a better (more robust) measure of spread.

c. The population is all SMSAs.

3.52

a. The quantile plots for homeownership rates for the years are given here.



b. The lower 10th percentile includes the 4 states (and DC) with the lowest ownership: DC, New York, Hawaii, California, and Rhode Island.

c. The lower 10th percentile states are the same in both 2002 and 1996. In 1985, swap Nevada for Rhode Island.

3.53

a.

Year	Mean	Median
1985	65.876	67.9
1996	66.843	68.2
2002	69.449	70.2

b. Since the homeownership percentages are skewed, the median is a more reliable measure of center.

c.

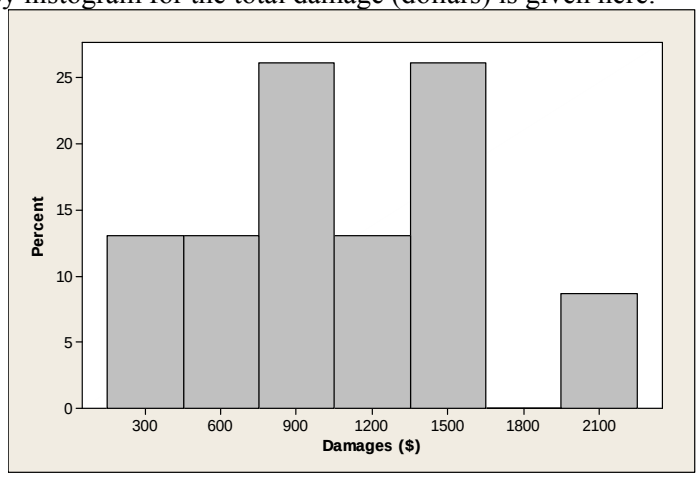
Year	StDev	MAD
1985	6.734	4.783
1996	6.688	4.697
2002	6.163	4.229

d. Skewed data is better served by a robust measure like MAD.

e. It appears home ownership rates are increasing and the variability is (slightly

3.54

a. A relative frequency histogram for the total damage (dollars) is given here.



b. Mean ≈ $1100

c. Mean = 25115/23 = $1091.96; Median = $1039

d. Because the mean is slightly larger than the median, it is likely that the distribution is slightly skewed to the right.

3.55

a. The means are given here:

Number of Members	Mean
1	93.75
2	98.652
3	113.3125
4	124.857
5+	131.90

b. Mean = 9024/83 = 108.7228

c. Yes, we can use the following method:
[20(93.75) + 23(98.652) + 16(113.3125) + 14(124.857) + 10(131.90)]/83
= 9023.994/83 = 108.7228

d. As the number of members increases, the mean expenditures tends to increase.

3.56

a. The standard deviations are given here:

Number of Members	StDev
1	37.99
2	20.19
3	33.62
4	44.70
5+	28.80

b. StDev = 35.42

c. Yes, we can use the following method:
[(20-1)(37.99) + (23-1)(20.19) + (16-1)(33.62) + (14-1)(44.70) + (10-1)(28.80)]/(83-1)
= 35.42

d. The group with the most variance contains 4 members.

3.57

a. The sample mean will be distorted by several large values which skew the distribution. State 5 and State 11 have more than 10 times as many plants destroyed as any other state; for arrests, States 1, 2, 8, and 12 exceed the other arrest figures substantially.

b. Plants: $\bar{y} = 10,166,919/15 = 677,794.60$

Arrests: $\bar{y} = 1425/15 = 95$

10% trimmed mean:

Plants: $\bar{y} = 1,565,604/11 = 142,327.64$

Arrests: $\bar{y} = 657/11 = 59.7$

20% trimmed mean:

Plants: $\bar{y} = 1,197,354/9 = 133,039.33$

Arrests: $\bar{y} = 372/9 = 41.30$

For plants, the 10% trimmed mean works well since it eliminates the effect of States 5 and 11. For arrests, the means differ because each takes some of the high values out of the calculation. It appears that the distribution is not skewed, but rather separated into at least two parts: states with high numbers of arrests and states with low numbers. By trimming the mean, we may be losing important information.

3.58

a. Mean = 2627.65/30 = \$87.59

b. $DJIA = \frac{2627.65}{0.155625} = 16844.5$

c. The DJIA does provide information about a population, the population of all companies. However, the sample is not a random sample.

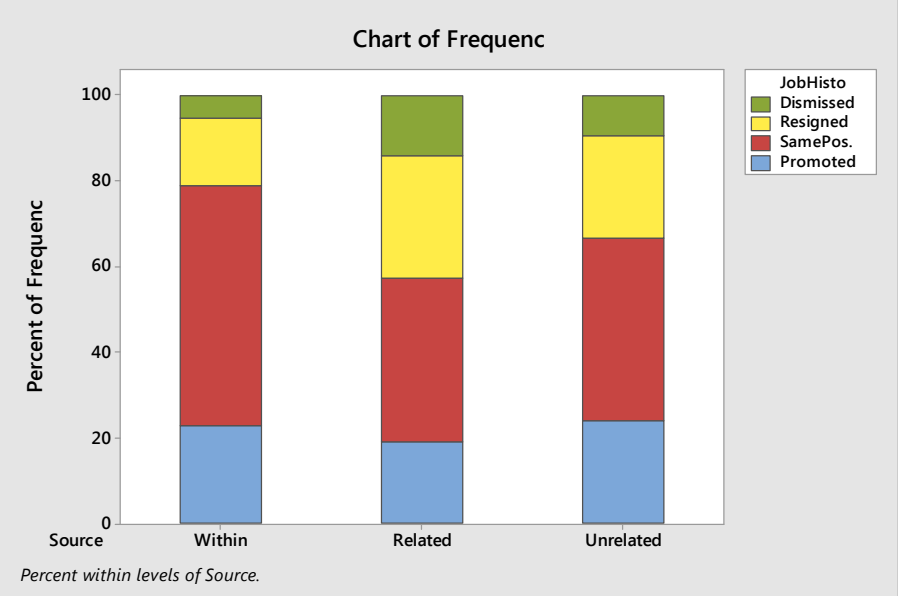
3.59

a. The job-history percentages within each source are given here:

	Source		
Job History	Within Firm	Related Business	Unrelated Business
Promoted	22.80	19.05	23.81
Same position	56.14	38.10	42.86
Resigned	15.80	28.57	23.81
Dismissed	5.26	14.28	9.52
Total	100 (n = 57)	100 (n = 21)	100 (n = 42)

b. If for each source we compute the percentages combined over the promoted and same position categories, we find that they are 78.94% for within firms, 57.15% for related business, and 66.67% for unrelated business. This ordering by source also holds for every job history category except the "promoted" one in which the three sources are nearly equal. It appears that a company does best when it selects its middle managers from within its own firm and worst when it takes its choices from a related firm.

c. The stacked bar chart gives us a visual representation of how hiring source may (or may not) be related to job trajectory. It appears hiring from within leads to a more successful middle manager.



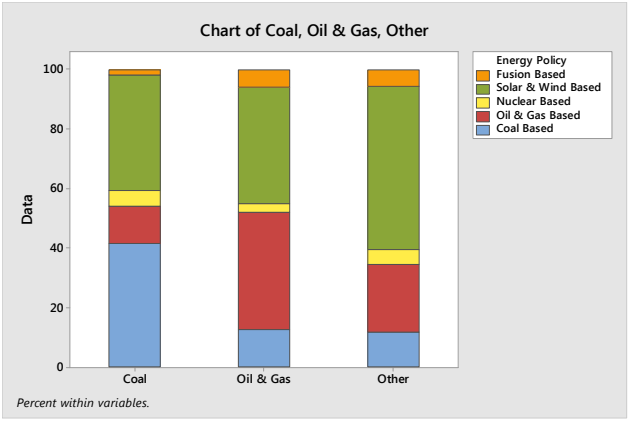
3.60

a. Proportions of each column

	Coal	Oil & Gas	Other	Total
Coal Based	0.413333	0.125	0.117778	0.175
Oil & Gas Based	0.126667	0.395	0.226667	0.25
Nuclear Based	0.053333	0.03	0.048889	0.045
Solar & Wind Based	0.386667	0.39	0.548889	0.47875
Fusion Based	0.02	0.06	0.057778	0.05125
Total	1	1	1	1

b. Yes. For both the coal and oil-gas states, the largest percentage of responses favored the type of energy produced in their own state.

c. The stacked bar chart shows Coal states are more likely to have Coal 'Energy policies, etc.



d. Solar and Wind Based has the highest support (47.8%) but it is important to note that most states don't fall into the 'Coal' or 'Oil & Gas' category (most in other) so there are (unsurprisingly) more people without a tie to a particular energy policy.

e. This question is open to interpretation. I'd venture that the general U.S. opinion is to desire more solar and wind power in theory, but the practice seems to belie this statement. Everyone 'prefers' a green energy policy, but not everyone 'practices what they preach'.

3.61 Arbitration seems to win the largest wage increases. If we assume that the Empirical Rule holds for these data, then a standard error for the mean of the arbitration figures would be $s/\sqrt{n} = 0.25$. Thus the mean increase after arbitration is (9.42 – 8.40)/0.25 ≈ 4 standard errors above the next largest mean, that