Chapter 3: Summarizing Data

3.1 a. Pie Chart should be plotted.

b. Bar Graph should be plotted.

3.2 a. A pie chart would not be appropriate since we are not proportionally allocating a sample or population into a number of categories.

b. Bar Graph

c. Yes. There was a decline from the late 80's to the early 90's. An abrupt surge upward occurred in 1994, 1995, and 1996.

3.3 Pie chart should be plotted.

3.4 Bar graph should be plotted.

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 for each value of Percentage of Use is probably the most effective plot. This plot would clearly demonstrate that the increase in number of lives saved by using a shoulder belt increased considerably as the percentage use increased.

3.6 a. Construct separate relative frequency histograms.

b. The histogram for the New Therapy has one more class than 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.

3.7 The plot has a bimodal shape. This would be an indication that there are two separate populations. However, the evidence is not very convincing because the individual plots were similar in shape with the exception that the New Therapy had a few times somewhat larger than the survival times obtained under the Standard Therapy.

3.8 a. The outlays have increased rapidly from 1980 to 1989, then dropped somewhat until 1991, when they increased again to 1992. The values had a slight decrease from 1992 through 1997.

b. The % GNP increased rapidly form 1980 to 1985, remain steady for a couple of years, then decreased fairly rapidly from 1987 through 1997 (except a minor jump from 1991 to 1992).

c. The two plots have very different trends. The public interest group's contention is not supported by either graph for the decade of the 90's.

3.9 a. There appears to have been a dramatic drop in verbal scores for both sexes from 1970 to 1980, with a greater drop in female scores. There appears to have been a slight drop in math scores from 1970 to 1980, then a slow steady rise from 1980 to 1996.

b. Yes. For the years 1967-1996, the difference between males and females has remained relatively constant.

c. The female verbal scores had a larger decrease for the years 1970-1980 than the male scores.

3.10 a. Relative frequency histograms for 1985 and 1996 should be plotted.

b. The two plots are very similar in shape. The proportion of states having a high percentage of homeownerships appears to have increased relative to 1985.

c. There was a very robust economy during this time period which may have allowed more people to purchase homes.

d. Congress would be able to determine that there has not been a large increase in homeownership during this period of a strong economy and decide to increase the tax deductions relative to homeowners.

3.11 Stem-and-Leaf Plot should be plotted.

3.12 The shapes of the 1985 and 1996 histograms and stem-and-leaf plots are asymmetric. The four plots are unimodal and left-skewed.

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 Mean = 243/16 = 15.1875 = 15.2, Median = (14+15)/2 = 14.5, Mode = 18

3.15 Mean = 283/16 = 17.6875 = 17.7, Median = (14+15)/2 = 14.5, Mode = 18

The median and mode are unalterered but the mean is inflated by the two large values.

3.16 10% of measurements = 10% of 16 = 1.6 $\approx$ 2. Thus, we delete the two largest and smallest data values. Since the 2 largest values are deleted in computing the 10% Trimmed Mean, its value is the same for both data sets:

10% Trimmed Mean = (11+11+12+13+14+14+15+17+17+18+18+18)/12 = 14.83. Thus, the extreme values do not have an effect. A 5% Trimmed Mean deletes only the smallest and largest values in the data set. Therefore, the second largest extreme value does enter into the calculation and hence would have an effect on the 5% Trimmed Mean. The effect would not be as dramatic as was seen in using the untrimmed mean.

3.17 a. Mean = 8.04, Median = 1.54

b. Terrestrial: Mean = 15.01, Median = 6.03
Aquatic: Mean = 0.38, Median = .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) results in a mean which is larger than 82% of the values in the data set.
Aquatic: Mean or median since the data set is relatively symmetric.

3.18 a. After removing the survival times of the two individuals who left the study, we obtain Mean = 46.3. The median can be calculated for all 11 patients, since we know that the values for the two indiduals who left the study were greater than the listed values of 57 and 60 which would place them in the upper half of the survival times. Thus, we obtain Median = 29.

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.

3.19 a. If we use all 14 failure times, we obtain Mean > 173.7 and Median $= 154$. In fact, we know the mean is greater than 173.7 since the failure times for two of engines are greater than the reported times of 300 hours.

b. The median would be unchanged if we replaced the failure times of 300 with the true failure times for the two engines that did not fail. However, the mean would be increased.

3.20 a. The values are given below:

Group	Mean	Median	Mode
1	2.923	2.805	no mode
2	1.592	1.565	1.55, 1.57
3	0.797	0.755	0.70

b. mean = 1.7707 median = 1.565 mode = 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 computed for each group, then the mean and median are both appropriate since the three groups have relatively symmetric distributions.

3.21 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 group have the same number of measurements, but is not true if the groups have different sample sizes. However, the average of the group modes and medians are different from the overall median and mode.

3.22 a. $\bar{y} = \sum y_i/n = \frac{40}{8} = 5$. Yes.

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. $s^2 = \frac{46}{8-1} = 6.57 \Rightarrow s = 2.56$.

3.23 a. s = 7.95

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

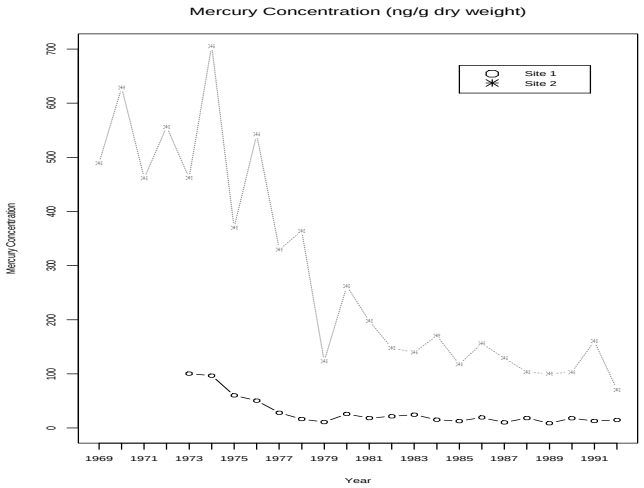
3.24 Age: $s \approx (10-2)/4 = 2$. The estimate is fairly close to the actual value of 2.56.

Experience: $s \approx (39-18)/4 = 5.25$. The estimate is somewhat smaller than the actual value of 7.95.

3.25 a. Luxury: $\bar{y} = 145.0, s = 27.6$
Budget: $\bar{y} = 46.1, s = 5.13$

b. 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.

3.26 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.

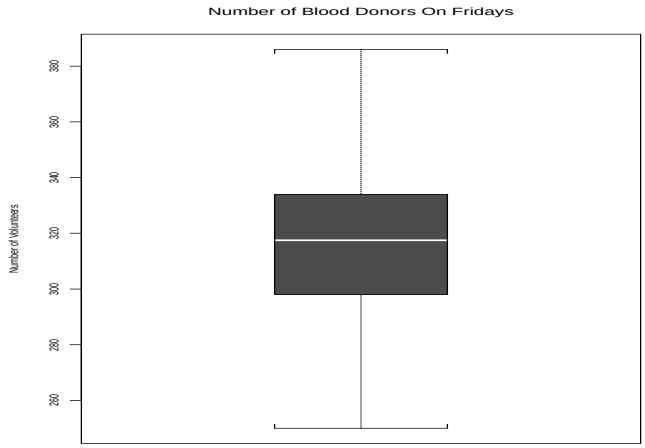
b. Site 1: Median = 18.25, Mean = 29.18
Site 2: Median = 184.1, Mean = 287.1
Both distributions are right skewed, thus the median is a more appropriate measure of center than is the mean. Site 1 has a considerably lower center than that of Site 2.

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

3.27 a. Stem-and-Leaf Plot is given here:

Stem	Leaf
2	5
2	677
2	9
3	00111
3	223333
3	5
3	67
3	8

b. Min=250, $Q_1 = (295 + 301)/2 = 298, Q_2 = (315 + 320)/2 = 317.5$, $Q_3 = (334 + 334)/2 = 334$, Max=386

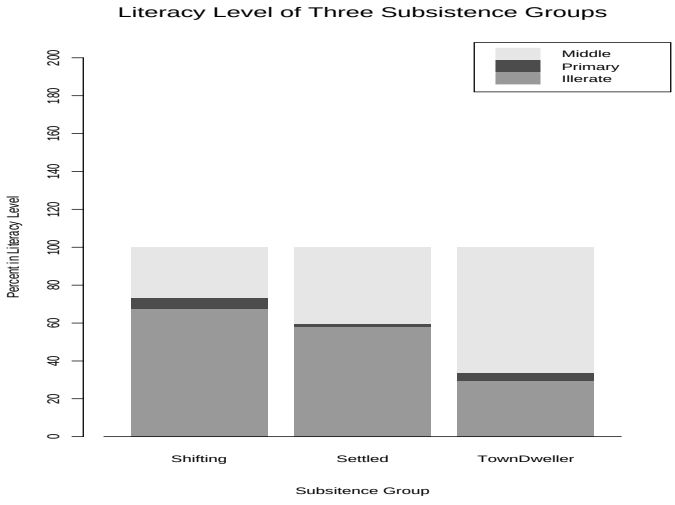


There are no outliers because $Q_1 - (1.5)IQR = 244 < 250$ and $Q_3 + (1.5)IQR = 388 > 386$. The distribution is approximately symmetric with no outliers.

3.28 a. CAN: $Q_1 \approx 1.45$, $Q_2 \approx 1.65$, $Q_3 \approx 2.4$
DRY: $Q_1 \approx 0.55$, $Q_2 \approx 0.60$, $Q_3 \approx 0.7$

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

3.29 a. Stacked Bar Graph is given here:



b. Illiterate: 46%, Primary Schooling: 4%, At Least Middle School: 50%

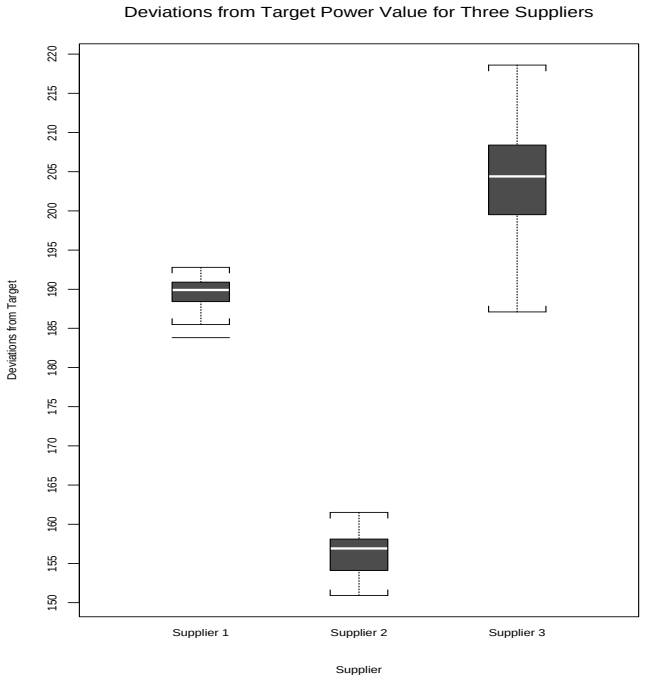
Shifting Cultivators: 27%, Settled Agriculturists: 21%, Town Dwellers: 51%

There is a marked difference in the distribution in 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.

3.30 a. The means and standard deviations are given here:

Supplier	$\bar{y}$	s
1	189.23	2.96
2	156.28	3.30
3	203.94	8.96

b. Side-by-side Boxplots are given here:

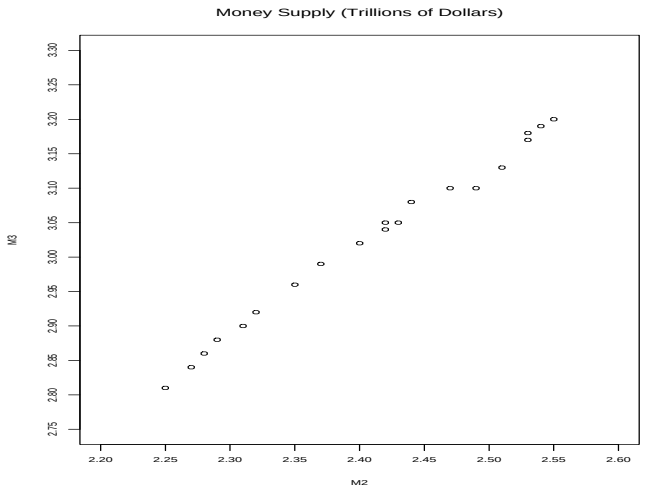


The three distributions are relatively symmetric but supplier 3 is considerably more variable and is shifted above supplier 1's values, which in turn are shifted above supplier 2's values.

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 deviates with essentially the same degree of variability as supplier 1.

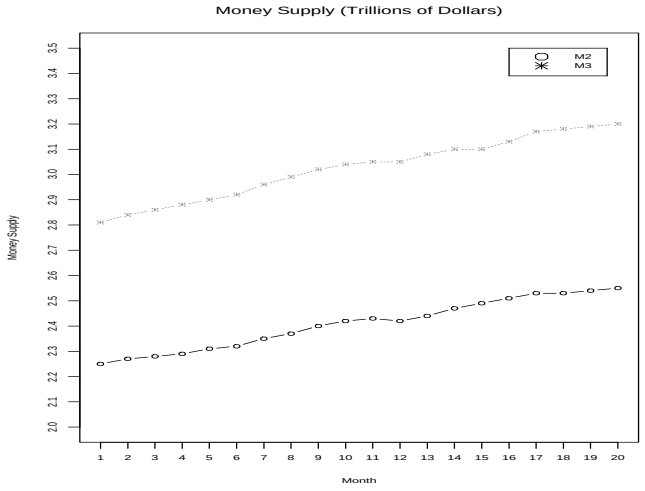
3.31 A scatterplot of M3 versus M2 is given here.



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 approximately increasing linear relationship.

3.32 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.

3.33 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.

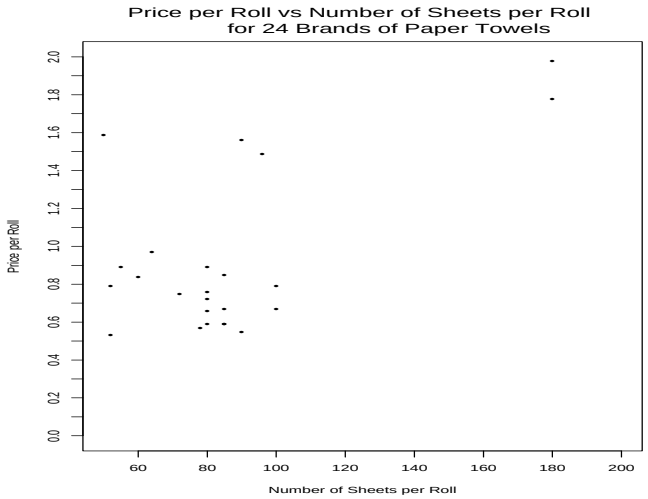
e. $\bar{y} \pm s \Rightarrow (-12.7, 127.7)$; yields 82%
$\bar{y} \pm 2s \Rightarrow (-82.9, 197.0)$; yields 94%
$\bar{y} \pm 3s \Rightarrow (-153.1, 268.1)$; yields 97%
These percentages do not match the Empirical Rule very well: 68%, 95%, and 99.7%

f. The Empirical Rule applies to data sets with roughly a "mound-shaped" histogram. The distribution of this data set is highly skewed right.

3.34 a. Price per roll: Mean = 0.9196, s = 0.4233
Price per sheet: Mean = 0.01091, s = 0.0059

b. Based on the standard deviation, the price per roll is more variable; but, if we look at the standard deviation as a percent of the sample mean, the price per sheet is more variable.

3.35 A scatterplot of is given here.



a. The points do not fall close to a straight line.

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 which will affect the price.

3.36 From the two boxplots, there are 5 unusual brands with regards to price per roll: \$1.49, \$1.56, \$1.59, \$1.78, and \$1.98. There are 2 unusual brands with respect to sheet per roll: 180 and 180.

3.37 a. Construct a relative frequency histogram

b. Highly skewed to the right

c. $1424 \pm 3488 \Rightarrow$ (-2063, 4912) contains $37/41 = 90.2\%$
$1424 \pm (2)3488 \Rightarrow$ (-5551, 8400) contains $38/41 = 92.7\%$
$1424 \pm (3)3488 \Rightarrow$ (-9039, 11888) contains $39/41 = 95.1\%$
These values do not match the percentages from the Empirical Rule: 68%, 95%, and 99.7%.

d. $1.48 \pm 1.54) \Rightarrow$ (-0.06, 3.02) contains $31/41 = 75.6\%$
$1.48 \pm (2)1.54 \Rightarrow$ (-1.60, 4.56) contains $40/41 = 97.6\%$
$1.48 \pm (3)1.54 \Rightarrow$ (-3.14, 6.10) contains $41/41 = 100\%$
These values closely match the percentages from the Empirical Rule: 68%, 95%, and 99.7%.

3.38 The values for the two years are given here:

Year	Mean	Median	Std. Dev.
1985	65.88	67.9	6.73
1996	66.84	68.2	6.69

a. The median is more appropriate since both distributions are left skewed.

b. There is not a large difference in the summary statistics between the two years.

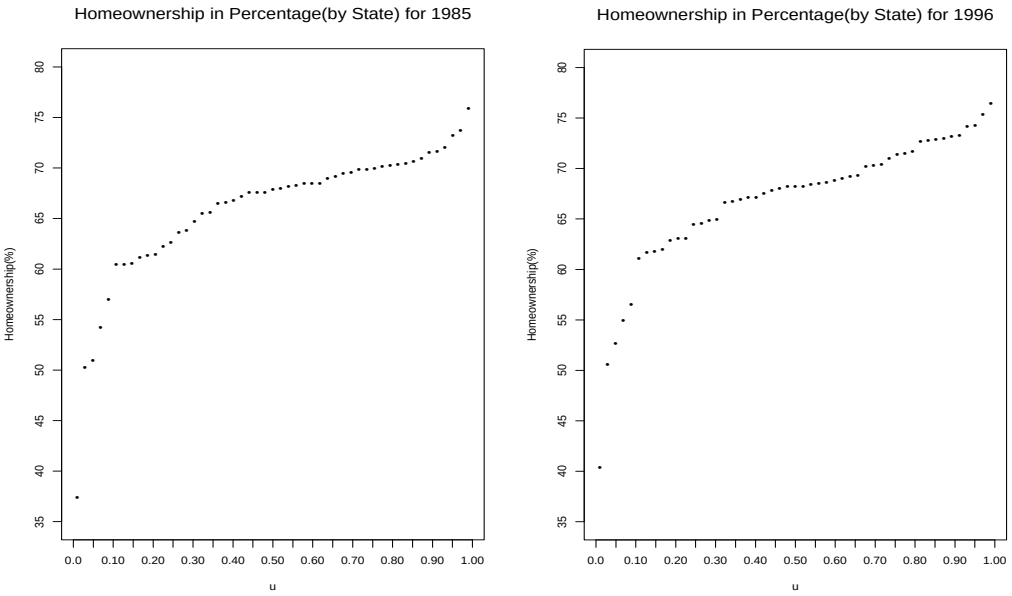
3.39 a. There has been very little change from 1985 to 1996.

b. Yes; For both 1985 and 1996, DC had extremely low ownership. New York and Hawaii had semi-low ownership.

c. No

d. The cost of homes is very high.

3.40 The quantile plots are given here.



a. The 20th percentile for 1996 is given by reading the vertical value on the graph for u = 0.20⇒ 63%. Thus, approximately 20% of the 1996 homeownership percentages are less than or equal to 63%.

b. The upper 10th percentile would correspond to the states having the largest 5 percentages: Michigan, Indiana, West Virginia, Minnesota, and Maine.

c. In 1985, the states falling in the upper 10th percentile are Pennsylvania, South Carolina, Wyoming, Maine, and West Virginia. There are only two states which fall into both groups.

3.41 a. Plot relative frequency histogram.

b. 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.

3.42 a. Mode = 688; yes, because it would indicate the most frequently occurring production.

b. Median = 688

c. Mean = 17664/26 = 679.58

d. Because the mean is somewhat less than the median, it is likely that the distribution skewed to the left.

3.43 The stem-and-leaf diagram is given here:

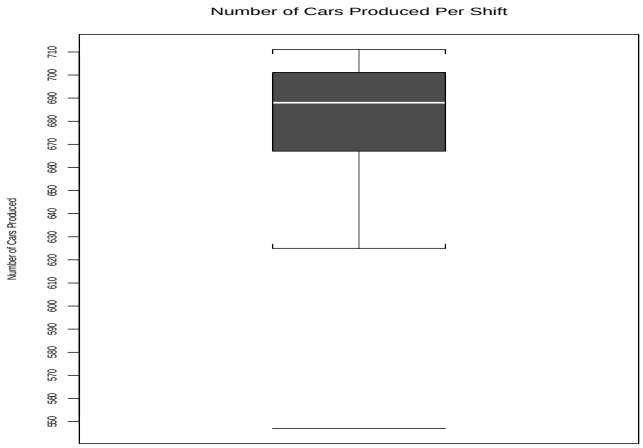
Stem	Leaf
54	7
55	
56	
57	
58	
59	
60	
61	
62	5
63	0
64	
65	6
66	4 7 7
67	79
68	8 88 88
69	1 4 7 9
70	0 1 2 333 8
71	1

Yes, the distribution is left skewed.

3.44 a. median(Q2) = 688, Ql = 667, Q3 = 701, IQR = 701-667 = 34

b. The inner fence boundaries are: 667-(1.5)(34) = 616 and 701+(1.5)(34) = 752. There is an outlier at 547.

c. The boxplot is given here:



3.45 a. New policy $\bar{y} = 2.27$, $s = 3.26$
Old policy: $\bar{y} = 4.60$, $s = 2.61$

b. Both the average number of sick days and the variation in number of sick days have decreased with new policy.

3.46 New policy: $\bar{y} = 1.93, \quad s = 2.31$

Old policy: $\bar{y} = 4.27, \quad s = 1.79$

Yes, the ranges are affected.

3.47 a. Average Price = 76.68

b. Range = 202.69

c. DJIA = 10856.2

d. Yes; The stocks covered on the NYSE.

No; The companies selected are the leading companies in the different sectors of business.

3.48 a. The job-history percentages within each source are give 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.

3.49 a. The value 62 reflects the number of respondents in coal producing states who preferred a national energy policy that encouraged coal production. The value 32.8 is of those who favored a coal policy, 32.80% came from major coal producing states. The value 41.3 tells us that 41.3% of those from coal states favored a coal policy. And the value 7.8 tells us that 7.8% of all the responses come from residents of major coal producing states who were in favor of a national energy policy that encouraged coal production.

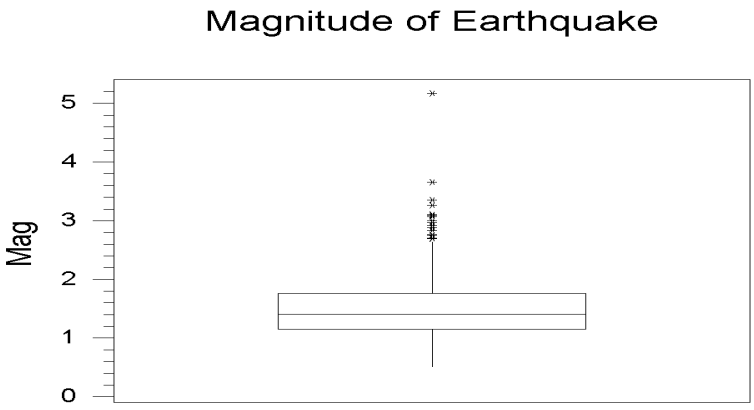
b. The column percentages because they displayed the distribution of opinions within each of the three types of states.

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

3.50 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 $s/\sqrt{n} = .25$. Thus the mean increase after arbitration is $(9.42 - 8.40)/.25 = 4$

standard errors above the next largest mean, that for Poststrike. Management, on the other hand, should favor negotiation. It has the smallest mean percentage wage increase and the smallest variance, or least risk.

3.51 a.



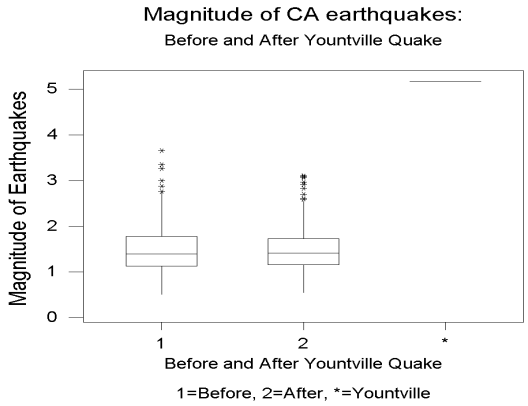
b. There are 28 bigger earthquakes with a median magnitude of 2.725.

c. A side-by-side boxplot and summary statistics for the magnitude of earthquakes are given here.

```
Descriptive Statistics: Magitude  by BeforeAfter

Variable                    N       Mean      Median     TrMean      StDev
Magitude   Before         230     1.4957      1.4000     1.4629     0.5545
           After          229     1.4955      1.4200     1.4691     0.5180
           Yountville       1     5.1700      5.1700     5.1700          *

Variable   BefAft      SE Mean    Minimum    Maximum         Q1         Q3
Magitude   Before       0.0366     0.5000     3.6600     1.1300     1.7800
           After        0.0342     0.5400     3.1100     1.1600     1.7350
           Yountville        *     5.1700     5.1700          *          *
```



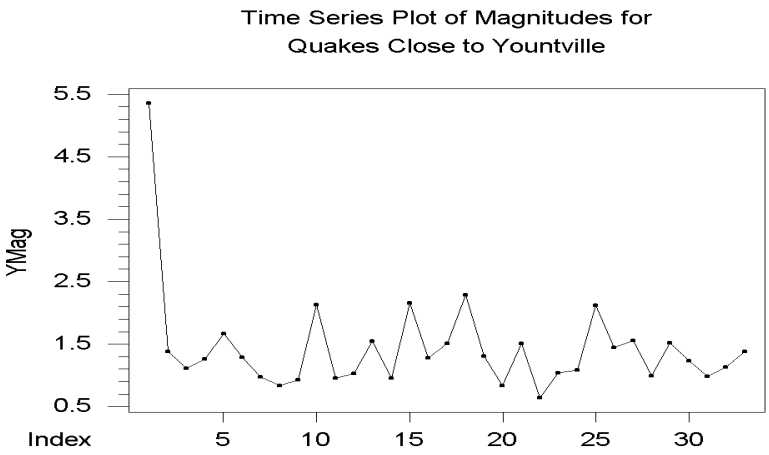
There does not seem to be an major differences in the magnitude of earthquakes before and after the Yountville quake.

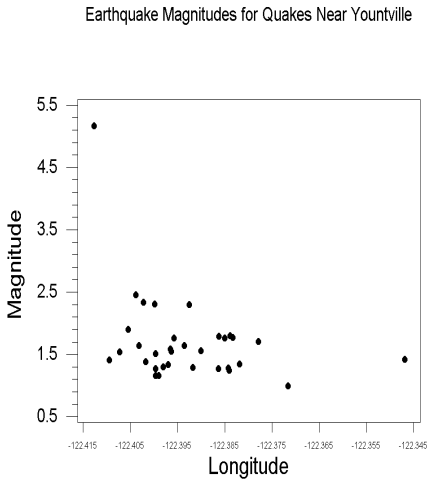
d. Summary statistics and plots are given here for earthquakes near Yountvile.

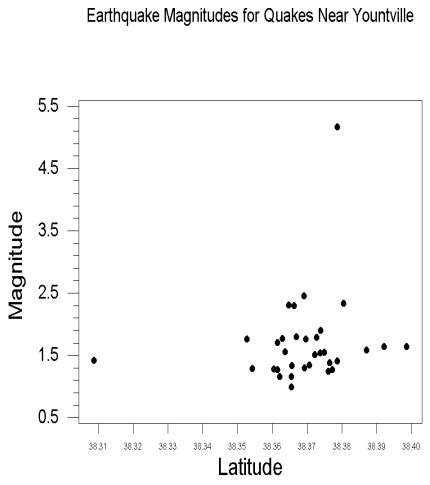
```
Descriptive Statistics: YLat, YLon, YMag

Variable             N       Mean     Median     TrMean      StDev    SE Mean
YLat                33     38.369     38.369     38.370      0.015      0.003
YLon                33    -122.39    -122.40    -122.39       0.01       0.00
YMag                33      1.694      1.550      1.591      0.721      0.125

Variable       Minimum    Maximum         Q1         Q3
YLat            38.309     38.399     38.363     38.376
YLon           -122.41    -122.35    -122.40    -122.38
YMag             0.990      5.170      1.295      1.780
```







There were no recorded earthquakes close to Yountville prior to the magnitude 5.17 quake but 33 after it occurred. The 5.17 magnitude quake was nearly centered in a north-south direction with respect to the "aftershocks", but the 5.17 quake was definitely shifted to the west of the "aftershocks".