CHAPTER 1

Section 1.1

1.

a. Houston Chronicle, Des Moines Register, Chicago Tribune, Washington Post

b. Capital One, Campbell Soup, Merrill Lynch, Pulitzer

c. Bill Jasper, Kay Reinke, Helen Ford, David Menedez

d. 1.78, 2.44, 3.5, 3.04

2.

a. 29.1 yd., 28.3 yd., 24.7 yd., 31.0 yd.

b. 432, 196, 184, 321

c. 2.1, 4.0, 3.2, 6.3

d. 0.07 g, 1.58 g, 7.1 g, 27.2 g

3.

a. In a sample of 100 VCRs, what are the chances that more than 20 need service while under warrantee? What are the chances than none need service while still under warrantee?

b. What proportion of all VCRs of this brand and model will need service within the warrantee period?

4.

a. Concrete: All living U.S. Citizens, all mutual funds marketed in the U.S., all books published in 1980.

Hypothetical: All grade point averages for University of California undergraduates during the next academic year. Page lengths for all books published during the next calendar year. Batting averages for all major league players during the next baseball season.

b. Concrete: Probability: In a sample of 5 mutual funds, what is the chance that all 5 have rates of return which exceeded 10% last year?

Statistics: If previous year rates-of-return for 5 mutual funds were 9.6, 14.5, 8.3, 9.9 and 10.2, can we conclude that the average rate for all funds was below 10%?

Conceptual: Probability: In a sample of 10 books to be published next year, how likely is it that the average number of pages for the 10 is between 200 and 250?

Statistics: If the sample average number of pages for 10 books is 227, can we be highly confident that the average for all books is between 200 and 245?

5.

a. No, the relevant conceptual population is all scores of all students who participate in the SI in conjunction with this particular statistics course.

b. The advantage to randomly choosing students to participate in the two groups is that we are more likely to get a sample representative of the population at large. If it were left to students to choose, there may be a division of abilities in the two groups which could unnecessarily affect the outcome of the experiment.

c. If all students were put in the treatment group there would be no results with which to compare the treatments.

6. One could take a simple random sample of students from all students in the California State University system and ask each student in the sample to report the distance form their hometown to campus. Alternatively, the sample could be generated by taking a stratified random sample by taking a simple random sample from each of the 23 campuses and again asking each student in the sample to report the distance from their hometown to campus. Certain problems might arise with self reporting of distances, such as recording error or poor recall. This study is enumerative because there exists a finite, identifiable population of objects from which to sample.

7. One could generate a simple random sample of all single family homes in the city or a stratified random sample by taking a simple random sample from each of the 10 district neighborhoods. From each of the homes in the sample the necessary variables would be collected. This would be an enumerative study because there exists a finite, identifiable population of objects from which to sample.

8.

a. Number observations equal 2 x 2 x 2 = 8

b. This could be called an analytic study because the data would be collected on an existing process. There is no sampling frame.

9.

a. There could be several explanations for the variability of the measurements. Among them could be measuring error, (due to mechanical or technical changes across measurements), recording error, differences in weather conditions at time of measurements, etc.

b. This could be called an analytic study because there is no sampling frame.

Section 1.2

10.

a. Minitab generates the following stem-and-leaf display of this data:

```
 5|9
 6|33588
 7|00234677889
 8|127
 9|077          stem: ones
10|7            leaf: tenths
11|368
```

What constitutes large or small variation usually depends on the application at hand, but an often-used rule of thumb is: the variation tends to be large whenever the spread of the data (the difference between the largest and smallest observations) is large compared to a representative value. Here, 'large' means that the percentage is closer to 100% than it is to 0%. For this data, the spread is 11 - 5 = 6, which constitutes 6/8 = .75, or, 75%, of the typical data value of 8. Most researchers would call this a large amount of variation.

b. The data display is not perfectly symmetric around some middle/representative value. There tends to be some positive skewness in this data.

c. In Chapter 1, outliers are data points that appear to be *very* different from the pack. Looking at the stem-and-leaf display in part (a), there appear to be no outliers in this data. (Chapter 2 gives a more precise definition of what constitutes an outlier).

d. From the stem-and-leaf display in part (a), there are 4 values greater than 10. Therefore, the proportion of data values that exceed 10 is 4/27 = .148, or, about 15%.

11.

6l	034	
6h	667899	
7l	00122244	
7h		Stem=Tens
8l	001111122344	Leaf=Ones
8h	5557899	
9l	03	
9h	58	

This display brings out the gap in the data:
There are no scores in the high 70's.

12. One method of denoting the pairs of stems having equal values is to denote the first stem by L, for 'low', and the second stem by H, for 'high'. Using this notation, the stem-and-leaf display would appear as follows:

3L	1	
3H	56678	
4L	000112222234	
4H	5667888	
5L	144	
5H	58	stem: tenths
6L	2	leaf: hundredths
6H	6678	
7L		
7H	5	

The stem-and-leaf display on the previous page shows that .45 is a good representative value for the data. In addition, the display is not symmetric and appears to be positively skewed. The spread of the data is .75 - .31 = .44, which is.44/.45 = .978, or about 98% of the typical value of .45. This constitutes a reasonably large amount of variation in the data. The data value .75 is a possible outlier

13.

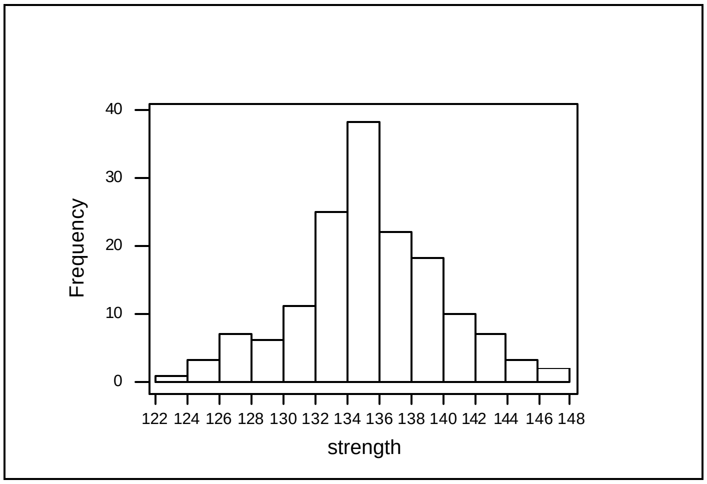
a.

12	2
12	445
12	6667777
12	889999
13	00011111111
13	2222222222333333333333333
13	44444444444444444455555555555555555555
13	6666666666667777777777
13	888888888888899999
14	0000001111
14	2333333
14	444
14	77

Leaf = ones
Stem = tens

The observations are highly concentrated at 134 – 135, where the display suggests the typical value falls.

b.



The histogram is symmetric and unimodal, with the point of symmetry at approximately 135.

14.

a.

2	23	stem units: 1.0
3	2344567789	leaf units: .10
4	01356889	
5	00001114455666789	
6	00001222233344456667789999	
7	00012233455555668	
8	02233448	
9	012233335666788	
10	2344455688	
11	2335999	
12	37	
13	8	
14	36	
15	0035	
16		
17		
18	9	

b. A representative value could be the median, 7.0.

c. The data appear to be highly concentrated, except for a few values on the positive side.

d. No, the data is skewed to the right, or positively skewed.

e. The value 18.9 appears to be an outlier, being more than two stem units from the previous value.

15.

Crunchy		Creamy
	2	2
644	3	69
77220	4	145
6320	5	3666
222	6	258
55	7	
0	8	

Both sets of scores are reasonably spread out. There appear to be no outliers. The three highest scores are for the crunchy peanut butter, the three lowest for the creamy peanut butter.

16.

a.

beams		cylinders
9	5	8
88533	6	16
98877643200	7	012488
721	8	13359
770	9	278
7	10	
863	11	2
	12	6
	13	
	14	1

The data appears to be slightly skewed to the right, or positively skewed. The value of 14.1 appears to be an outlier. Three out of the twenty, 3/20 or .15 of the observations exceed 10 Mpa.

b. The majority of observations are between 5 and 9 Mpa for both beams and cylinders, with the modal class in the 7 Mpa range. The observations for cylinders are more variable, or spread out, and the maximum value of the cylinder observations is higher.

c. Dot Plot

```
          . .  .  :..  : .: . . .   :         .        .         .
          -+---------+---------+---------+---------+---------+-----
cylinder
         6.0       7.5       9.0      10.5      12.0      13.5
```

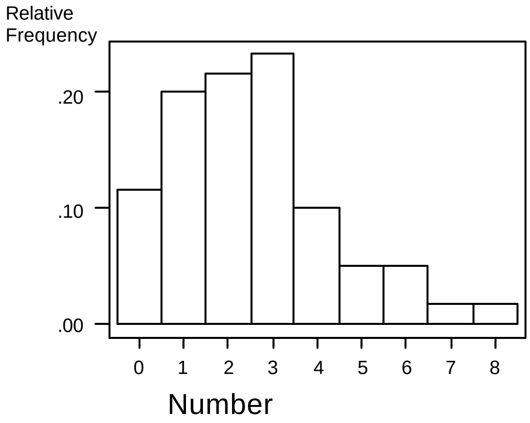
17.

a.

Number Nonconforming	Frequency	RelativeFrequency(Freq/60)
0	7	0.117
1	12	0.200
2	13	0.217
3	14	0.233
4	6	0.100
5	3	0.050
6	3	0.050
7	1	0.017
8	1	0.017
doesn't add exactly to 1 because relative frequencies have been rounded		1.001

b. The number of batches with at most 5 nonconforming items is 7+12+13+14+6+3 = 55, which is a proportion of 55/60 = .917. The proportion of batches with (strictly) fewer than 5 nonconforming items is 52/60 = .867. Notice that these proportions could also have been computed by using the relative frequencies: e.g., proportion of batches with 5 or fewer nonconforming items = 1- (.05+.017+.017) = .916; proportion of batches with fewer than 5 nonconforming items = 1 - (.05+.05+.017+.017) = .866.

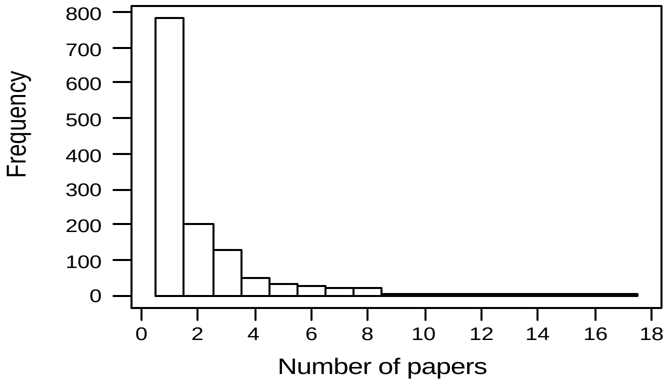
c. The following is a Minitab histogram of this data. The center of the histogram is somewhere around 2 or 3 and it shows that there is some positive skewness in the data. Using the rule of thumb in Exercise 1, the histogram also shows that there is a lot of spread/variation in this data.



18.

a.

The following histogram was constructed using Minitab:



The most interesting feature of the histogram is the heavy positive skewness of the data.

Note: One way to have Minitab automatically construct a histogram from grouped data such as this is to use Minitab's ability to enter multiple copies of the same number by typing, for example, 784(1) to enter 784 copies of the number 1. The frequency data in this exercise was entered using the following Minitab commands:

```
MTB > set c1
DATA> 784(1) 204(2) 127(3) 50(4) 33(5) 28(6) 19(7) 19(8)
DATA> 6(9) 7(10) 6(11) 7(12) 4(13) 4(14) 5(15) 3(16) 3(17)
DATA> end
```

b. From the frequency distribution (or from the histogram), the number of authors who published at least 5 papers is 33+28+19+…+5+3+3 = 144, so the proportion who published 5 or more papers is 144/1309 = .11, or 11%. Similarly, by adding frequencies and dividing by n = 1309, the proportion who published 10 or more papers is 39/1309 = .0298, or about 3%. The proportion who published more than 10 papers (i.e., 11 or more) is 32/1309 = .0245, or about 2.5%.

c. No. Strictly speaking, the class described by ' ≥15 ' has no upper boundary, so it is impossible to draw a rectangle above it having finite area (i.e., frequency).

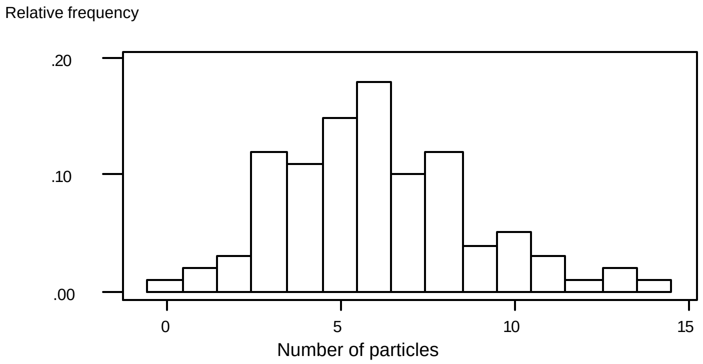
d. The category 15-17 does have a finite width of 2, so the cumulated frequency of 11 can be plotted as a rectangle of height 6.5 over this interval. The basic rule is to make the area of the bar equal to the class frequency, so area = 11 = (width)(height) = 2(height) yields a height of 6.5.

19.

a. From this frequency distribution, the proportion of wafers that contained at least one particle is (100-1)/100 = .99, or 99%. Note that it is much easier to subtract 1 (which is the number of wafers that contain 0 particles) from 100 than it would be to add all the frequencies for 1, 2, 3,… particles. In a similar fashion, the proportion containing at least 5 particles is (100 - 1-2-3-12-11)/100 = 71/100 = .71, or, 71%.

b. The proportion containing between 5 and 10 particles is (15+18+10+12+4+5)/100 = 64/100 = .64, or 64%. The proportion that contain strictly between 5 and 10 (meaning strictly *more* than 5 and strictly *less* than 10) is (18+10+12+4)/100 = 44/100 = .44, or 44%.

c. The following histogram was constructed using Minitab. The data was entered using the same technique mentioned in the answer to exercise 8(a). The histogram is *almost* symmetric and unimodal; however, it has a few relative maxima (i.e., modes) and has a very slight positive skew.



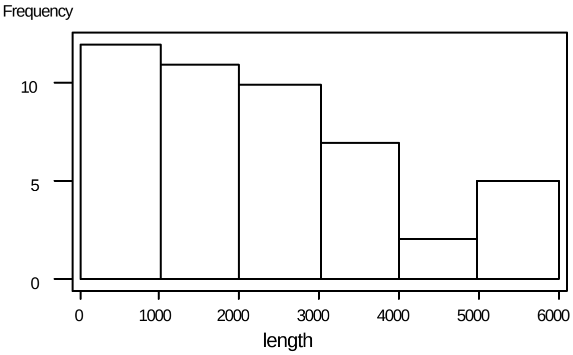
20.

a. The following stem-and-leaf display was constructed:

0	123334555599	
1	00122234688	stem: thousands
2	1112344477	leaf: hundreds
3	0113338	
4	37	
5	23778	

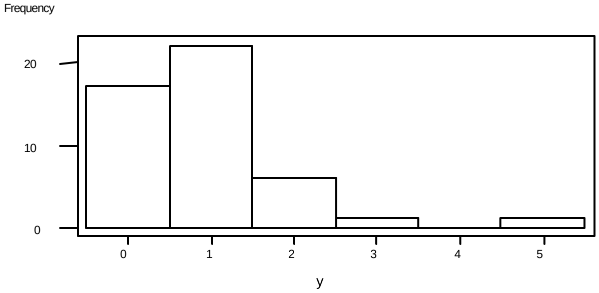
A typical data value is somewhere in the low 2000's. The display is almost unimodal (the stem at 5 would be considered a mode, the stem at 0 another) and has a positive skew.

b. A histogram of this data, using classes of width 1000 centered at 0, 1000, 2000, 6000 is shown below. The proportion of subdivisions with total length less than 2000 is (12+11)/47 = .489, or 48.9%. Between 200 and 4000, the proportion is (7 + 2)/47 = .191, or 19.1%. The histogram shows the same general shape as depicted by the stem-and-leaf in part (a).

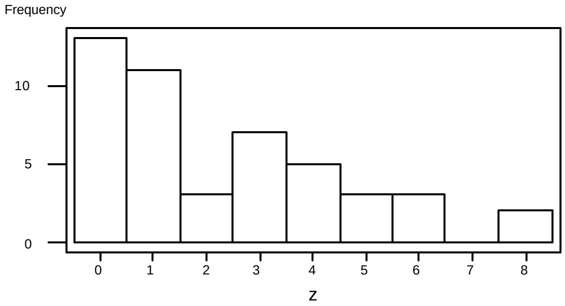


21.

a. A histogram of the y data appears below. From this histogram, the number of subdivisions having no cul-de-sacs (i.e., y = 0) is 17/47 = .362, or 36.2%. The proportion having at least one cul-de-sac (y ≥ 1) is (47-17)/47 = 30/47 = .638, or 63.8%. Note that subtracting the number of cul-de-sacs with y = 0 from the total, 47, is an easy way to find the number of subdivisions with y ≥ 1.



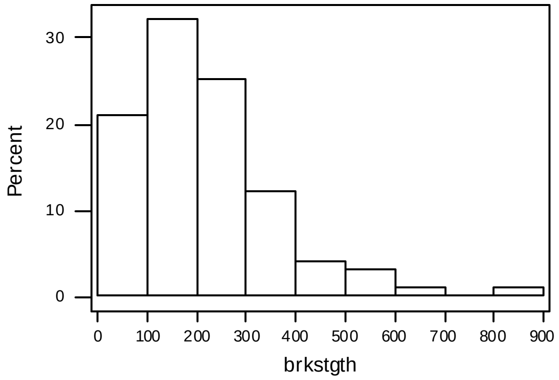
b. A histogram of the z data appears below. From this histogram, the number of subdivisions with at most 5 intersections (i.e., z ≤ 5) is 42/47 = .894, or 89.4%. The proportion having fewer than 5 intersections (z < 5) is 39/47 = .830, or 83.0%.



22. A very large percentage of the data values are greater than 0, which indicates that most, but not all, runners do slow down at the end of the race. The histogram is also positively skewed, which means that some runners slow down a *lot* compared to the others. A typical value for this data would be in the neighborhood of 200 seconds. The proportion of the runners who ran the last 5 km faster than they did the first 5 km is very small, about 1% or so.

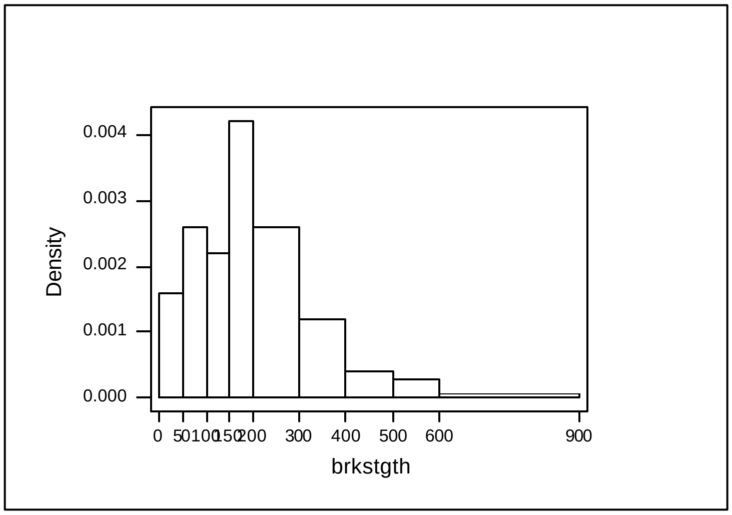
23.

a.



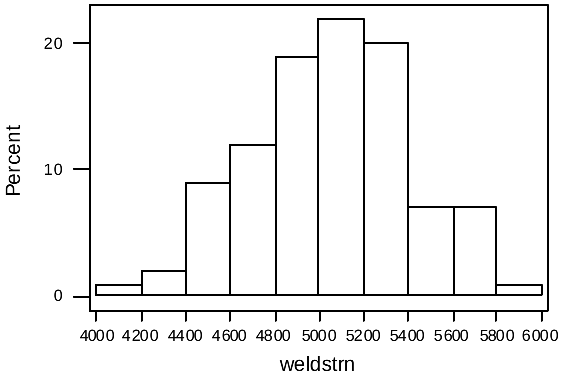
The histogram is skewed right, with a majority of observations between 0 and 300 cycles. The class holding the most observations is between 100 and 200 cycles.

b.

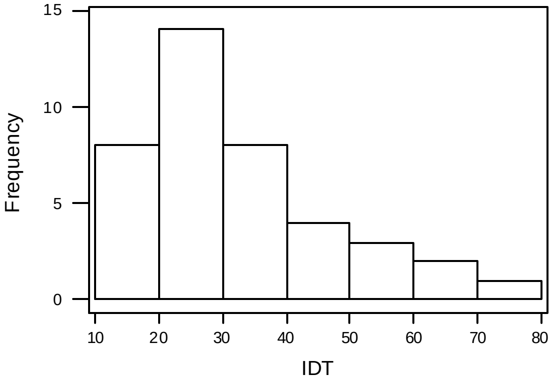


c [proportion ≥ 100] = 1 – [proportion < 100] = 1 - .21 = .79

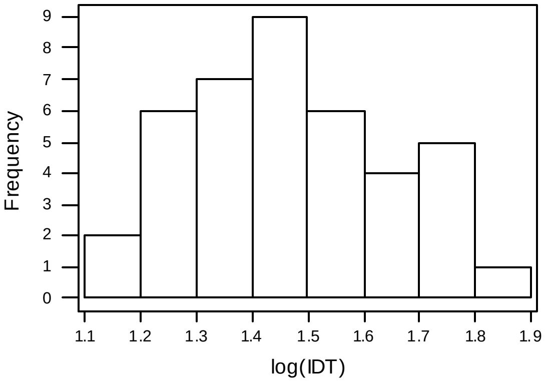
24.



25. Histogram of original data:



Histogram of transformed data:

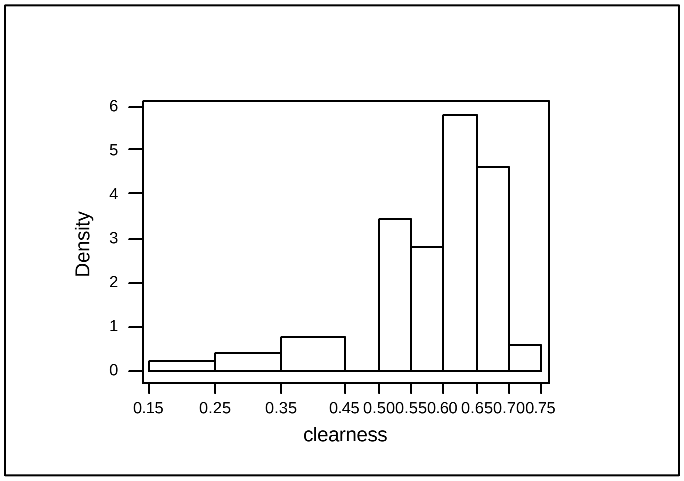


The transformation creates a much more symmetric, mound-shaped histogram.

26.

a.

Class Intervals	**Frequency**	**Rel. Freq.**
.15 -< .25	8	0.02192
.25 -< .35	14	0.03836
.35 -< .45	28	0.07671
.45 -< .50	24	0.06575
.50 -< .55	39	0.10685
.55 -< .60	51	0.13973
.60 -< .65	106	0.29041
.65 -< .70	84	0.23014
.70 -< .75	11	0.03014
	n=365	1.00001



b. The proportion of days with a clearness index smaller than .35 is $\frac{(8+4)}{365} = .06$, or 6%.

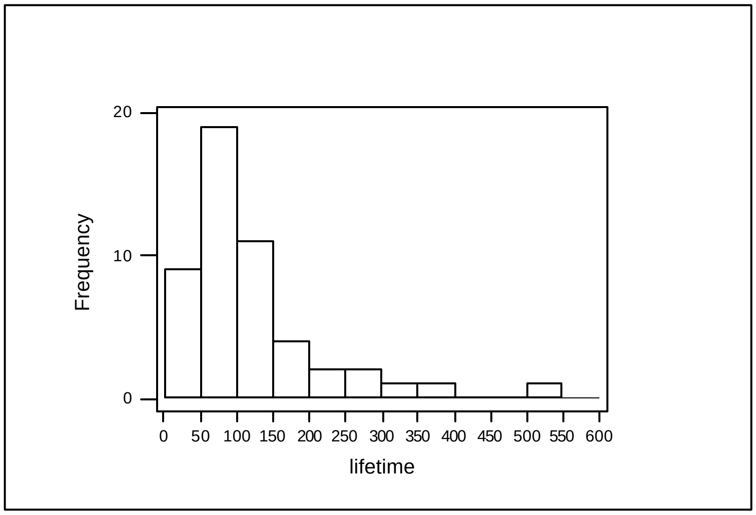
c. The proportion of days with a clearness index of at least .65 is $\frac{(84+11)}{365} = .26$, or 26%.

27.

a. The endpoints of the class intervals overlap. For example, the value 50 falls in both of the intervals '0 – 50' and '50 – 100'.

b.

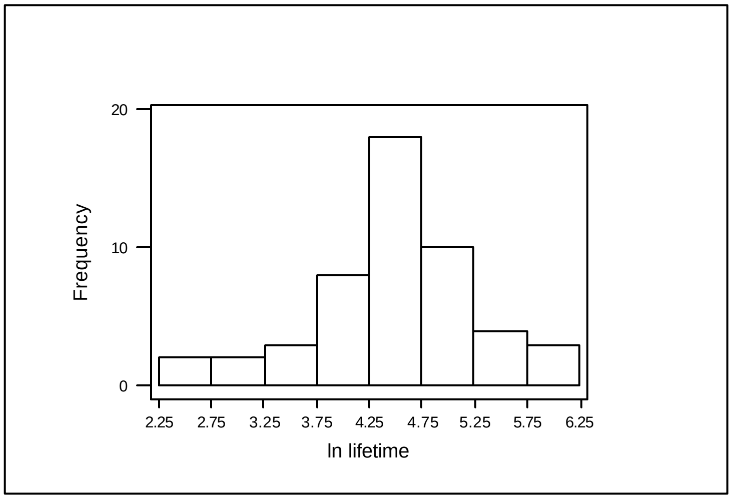
Class Interval	Frequency	Relative Frequency
0 - < 50	9	0.18
50 - < 100	19	0.38
100 - < 150	11	0.22
150 - < 200	4	0.08
200 - < 250	2	0.04
250 - < 300	2	0.04
300 - < 350	1	0.02
350 - < 400	1	0.02
>= 400	1	0.02
	50	1.00



The distribution is skewed to the right, or positively skewed. There is a gap in the histogram, and what appears to be an outlier in the '500 – 550' interval.

c.

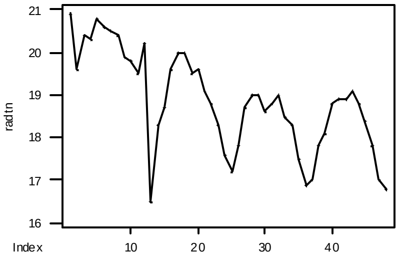
Class Interval	Frequency	Relative Frequency
2.25 - < 2.75	2	0.04
2.75 - < 3.25	2	0.04
3.25 - < 3.75	3	0.06
3.75 - < 4.25	8	0.16
4.25 - < 4.75	18	0.36
4.75 - < 5.25	10	0.20
5.25 - < 5.75	4	0.08
5.75 - < 6.25	3	0.06



The distribution of the natural logs of the original data is much more symmetric than the original.

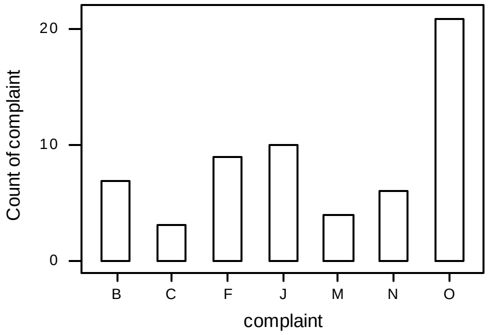
d. The proportion of lifetime observations in this sample that are less than 100 is .18 + .38 = .56, and the proportion that is at least 200 is .04 + .04 + .02 + .02 + .02 = .14.

28. There are seasonal trends with lows and highs 12 months apart.

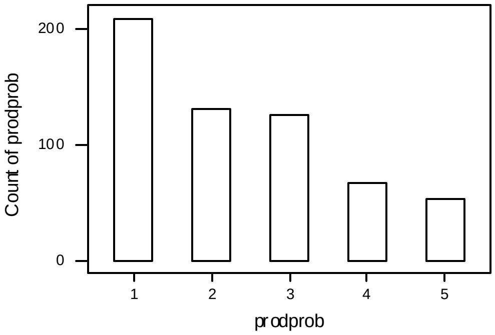


29.

Complaint	Frequency	Relative Frequency
B	7	0.1167
C	3	0.0500
F	9	0.1500
J	10	0.1667
M	4	0.0667
N	6	0.1000
O	21	0.3500
	60	1.0000



30.



1. incorrect comp onent
2. missing component
3. failed component
4. insufficient solder
5. excess solder

31.

Class	Frequency	Relative Frequency	Cumulative Relative Frequency
0.0 - under 4.0	2	2	0.050
4.0 - under 8.0	14	16	0.400
8.0 - under 12.0	11	27	0.675
12.0 - under 16.0	8	35	0.875
16.0 - under 20.0	4	39	0.975
20.0 - under 24.0	0	39	0.975
24.0 - under 28.0	1	40	1.000

32.

a. The frequency distribution is:

Class	Relative Frequency	Class	Relative Frequency
0-< 150	.193	900-<1050	.019
150-< 300	.183	1050-<1200	.029
300-< 450	.251	1200-<1350	.005
450-< 600	.148	1350-<1500	.004
600-< 750	.097	1500-<1650	.001
750-< 900	.066	1650-<1800	.002
		1800-<1950	.002

The relative frequency distribution is almost unimodal and exhibits a large positive skew. The typical middle value is somewhere between 400 and 450, although the skewness makes it difficult to pinpoint more exactly than this.

b. The proportion of the fire loads less than 600 is .193+.183+.251+.148 = .775. The proportion of loads that are at least 1200 is .005+.004+.001+.002+.002 = .014.

c. The proportion of loads between 600 and 1200 is 1 - .775 - .014 = .211.

Section 1.3

33.

a. $\bar{x} = 192.57$, $\tilde{x} = 189$. The mean is larger than the median, but they are still fairly close together.

b. Changing the one value, $\bar{x} = 189.71$, $\tilde{x} = 189$. The mean is lowered, the median stays the same.

c. $\bar{x}_{tr} = 191.0$. $\frac{1}{14} = .07$ or 7% trimmed from each tail.

d. For n = 13, Σx = (119.7692) x 13 = 1,557
For n = 14, Σx = 1,557 + 159 = 1,716

$$\bar{x} = \frac{1716}{14} = 122.5714 \text{ or } 122.6$$

34.

a. The sum of the n = 11 data points is 514.90, so $\bar{x}$ = 514.90/11 = 46.81.

b. The sample size (n = 11) is odd, so there will be a middle value. Sorting from smallest to largest: 4.4 16.4 22.2 30.0 33.1 36.6 40.4 66.7 73.7 81.5 109.9. The sixth value, 36.6 is the middle, or median, value. The mean differs from the median because the largest sample observations are much further from the median than are the smallest values.

c. Deleting the smallest (x = 4.4) and largest (x = 109.9) values, the sum of the remaining 9 observations is 400.6. The trimmed mean $\bar{x}_{tr}$ is 400.6/9 = 44.51. The trimming percentage is 100(1/11) ≈ 9.1%. $\bar{x}_{tr}$ lies between the mean and median.

35.

a. The sample mean is $\bar{x}$ = (100.4/8) = 12.55.

The sample size (n = 8) is even. Therefore, the sample median is the average of the (n/2) and (n/2) + 1 values. By sorting the 8 values in order, from smallest to largest: 8.0 8.9 11.0 12.0 13.0 14.5 15.0 18.0, the forth and fifth values are 12 and 13. The sample median is (12.0 + 13.0)/2 = 12.5.

The 12.5% trimmed mean requires that we first trim (.125)(n) or 1 value from the ends of the ordered data set. Then we average the remaining 6 values. The 12.5% trimmed mean $\bar{x}_{tr(12.5)}$ is 74.4/6 = 12.4.

All three measures of center are similar, indicating little skewness to the data set.

b. The smallest value (8.0) could be increased to any number below 12.0 (a change of less than 4.0) without affecting the value of the sample median.

c. The values obtained in part (a) can be used directly. For example, the sample mean of 12.55 psi could be re-expressed as

$$(12.55 \text{ psi}) \text{ x} \left(\frac{1ksi}{2.2\,psi}\right) = 5.70\,ksi.$$

36.

a. A stem-and leaf display of this data appears below:

32	55	stem: ones
33	49	leaf: tenths
34		
35	6699	
36	34469	
37	03345	
38	9	
39	2347	
40	23	
41		
42	4	

The display is reasonably symmetric, so the mean and median will be close.

b. The sample mean is $\bar{x}$ = 9638/26 = 370.7. The sample median is $\tilde{x}$ = (369+370)/2 = 369.50.

c. The largest value (currently 424) could be increased by any amount. Doing so will not change the fact that the middle two observations are 369 and 170, and hence, the median will not change. However, the value x = 424 can not be changed to a number less than 370 (a change of 424-370 = 54) since that *will* lower the values(s) of the two middle observations.

d. Expressed in minutes, the mean is (370.7 sec)/(60 sec) = 6.18 min; the median is 6.16 min.

37. $\bar{x} = 12.01$, $\tilde{x} = 11.35$, $\bar{x}_{tr(10)} = 11.46$. The median or the trimmed mean would be good choices because of the outlier 21.9.

38.

a. The reported values are (in increasing order) 110, 115, 120, 120, 125, 130, 130, 135, and 140. Thus the median of the reported values is 125.

b. 127.6 is reported as 130, so the median is now 130, a very substantial change. When there is rounding or grouping, the median can be highly sensitive to small change.

39.

a. $\Sigma x_l = 16.475$ so $\bar{x} = \dfrac{16.475}{16} = 1.0297$

$$\tilde{x} = \frac{(1.007 + 1.011)}{2} = 1.009$$

b. 1.394 can be decreased until it reaches 1.011(the largest of the 2 middle values) – i.e. by 1.394 – 1.011 = .383, If it is decreased by more than .383, the median will change.

40. $\tilde{x} = 60.8$

$\bar{x}_{tr(25)} = 59.3083$

$\bar{x}_{tr(10)} = 58.3475$

$\bar{x} = 58.54$

All four measures of center have about the same value.

41.

a. $7/10 = .70$

b. $\bar{x} = .70$ = proportion of successes

c. $\dfrac{s}{25} = .80$ so s = (0.80)(25) = 20

total of 20 successes

20 – 7 = 13 of the new cars would have to be successes

42.

a. $\bar{y} = \dfrac{\Sigma y_i}{n} = \dfrac{\Sigma(x_i + c)}{n} = \dfrac{\Sigma x_i}{n} + \dfrac{nc}{n} = \bar{x} + c$

$\tilde{y}$ = the median of $(x_1 + c, x_2 + c, ..., x_n + c)$ = median of $(x_1, x_2, ..., x_n) + c = \tilde{x} + c$

b. $\bar{y} = \dfrac{\Sigma y_i}{n} = \dfrac{\Sigma(x_i \cdot c)}{n} = \dfrac{c\Sigma x_i}{n} = c\bar{x}$

$\tilde{y} = (cx_1, cx_2, ..., cx_n) = c \cdot median(x_1, x_2, ..., x_n) = c\tilde{x}$

43. median = $\dfrac{(57 + 79)}{2} = 68.0$, 20% trimmed mean = 66.2, 30% trimmed mean = 67.5.

Section 1.4

44.

a. range = 49.3 – 23.5 = 25.8

b.

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	x_i^2
29.5	-1.53	2.3409	870.25
49.3	18.27	333.7929	2430.49
30.6	-0.43	0.1849	936.36
28.2	-2.83	8.0089	795.24
28.0	-3.03	9.1809	784.00
26.3	-4.73	22.3729	691.69
33.9	2.87	8.2369	1149.21
29.4	-1.63	2.6569	864.36
23.5	-7.53	56.7009	552.25
31.6	0.57	0.3249	998.56

$\Sigma x = 310.3$ $\quad \Sigma(x_i - \bar{x}) = 0$ $\quad \Sigma(x_i - \bar{x})^2 = 443.801$ $\quad \Sigma(x_i^2) = 10{,}072.41$

$\bar{x} = 31.03$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{443.801}{9} = 49.3112$$

c. $s = \sqrt{s^2} = 7.0222$

d. $s^2 = \dfrac{\Sigma x^2 - (\Sigma x)^2 / n}{n-1} = \dfrac{10{,}072.41 - (310.3)^2 / 10}{9} = 49.3112$

45.

a. $\bar{x} = \frac{1}{n}\sum_i x_i$ = 577.9/5 = 115.58. Deviations from the mean:
116.4 - 115.58 = .82, 115.9 - 115.58 = .32, 114.6 -115.58 = -.98,
115.2 - 115.58 = -.38, and 115.8-115.58 = .22.

b. $s^2 = [(.82)^2 + (.32)^2 + (-.98)^2 + (-.38)^2 + (.22)^2]/(5-1)$ = 1.928/4 =.482,
so s = .694.

c. $\sum_i x_i^2$ = 66,795.61, so $s^2 = \frac{1}{n-1}\left[\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2\right] =$
$[66{,}795.61 - (577.9)^2/5]/4$ = 1.928/4 = .482.

d. Subtracting 100 from all values gives $\bar{x} = 15.58$, all deviations are the same as in part b, and the transformed variance is identical to that of part b.

46.

a. $\bar{x} = \frac{1}{n}\sum_i x_i$ = 14438/5 = 2887.6. The sorted data is: 2781 2856 2888 2900 3013, so the sample median is $\tilde{x}$ = 2888.

b. Subtracting a constant from each observation shifts the data, but does not change its sample variance (Exercise 16). For example, by subtracting 2700 from each observation we get the values 81, 200, 313, 156, and 188, which are smaller (fewer digits) and easier to work with. The sum of squares of this transformed data is 204210 and its sum is 938, so the computational formula for the variance gives s^2 = [204210-(938)^2/5]/(5-1) = 7060.3.

47. The sample mean, $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}(1{,}162) = \bar{x} = 116.2$.

The sample standard deviation, $s = \sqrt{\dfrac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\dfrac{140{,}992 - \frac{(1{,}162)^2}{10}}{9}} = 25.75$

On average, we would expect a fracture strength of 116.2. In general, the size of a typical deviation from the sample mean (116.2) is about 25.75. Some observations may deviate from 116.2 by more than this and some by less.

48. Using the computational formula, $s^2 = \frac{1}{n-1}\left[\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2\right]$ =

[3,587,566-(9638)^2/26]/(26-1) = 593.3415, so s = 24.36. In general, the size of a typical deviation from the sample mean (370.7) is about 24.4. Some observations may deviate from 370.7 by a little more than this, some by less.

49.

a. $\Sigma x = 2.75 + ... + 3.01 = 56.80$, $\Sigma x^2 = (2.75)^2 + ... + (3.01)^2 = 197.8040$

b. $s^2 = \dfrac{197.8040 - (56.80)^2/17}{16} = \dfrac{8.0252}{16} = .5016, \;\; s = .708$

50. First, we need $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{27}(20{,}179) = 747.37$. Then we need the sample standard deviation $s = \sqrt{\frac{24{,}657{,}511 - \frac{(20{,}179)^2}{27}}{26}} = 606.89$. The maximum award should be $\bar{x} + 2s = 747.37 + 2(606.89) = 1961.16$, or in dollar units, $1,961,160. This is quite a bit less than the $3.5 million that was awarded originally.

51.

a. $\Sigma x = 2563$ and $\Sigma x^2 = 368{,}501$, so

$$s^2 = \frac{[368{,}501 - (2563)^2/19]}{18} = 1264.766 \text{ and } s = 35.564$$

b. If y = time in minutes, then y = cx where $c = \frac{1}{60}$, so

$$s_y^2 = c^2 s_x^2 = \frac{1264.766}{3600} = .351 \text{ and } s_y = cs_x = \frac{35.564}{60} = .593$$

52. Let *d* denote the fifth deviation. Then $.3 + .9 + 1.0 + 1.3 + d = 0$ or $3.5 + d = 0$, so $d = -3.5$. One sample for which these are the deviations is $x_1 = 3.8,\ x_2 = 4.4,$ $x_3 = 4.5,\ x_4 = 4.8,\ x_5 = 0.$ (obtained by adding 3.5 to each deviation; adding any other number will produce a different sample with the desired property)

53.

a. lower half: 2.34 2.43 2.62 2.74 2.74 2.75 2.78 3.01 3.46
upper half: 3.46 3.56 3.65 3.85 3.88 3.93 4.21 4.33 4.52
Thus the lower fourth is 2.74 and the upper fourth is 3.88.

b. $f_s = 3.88 - 2.74 = 1.14$

c. f_s wouldn't change, since increasing the two largest values does not affect the upper fourth.

d. By at most .40 (that is, to anything not exceeding 2.74), since then it will not change the lower fourth.

e. Since n is now even, the lower half consists of the smallest 9 observations and the upper half consists of the largest 9. With the lower fourth = 2.74 and the upper fourth = 3.93, $f_s = 1.19$.

54.

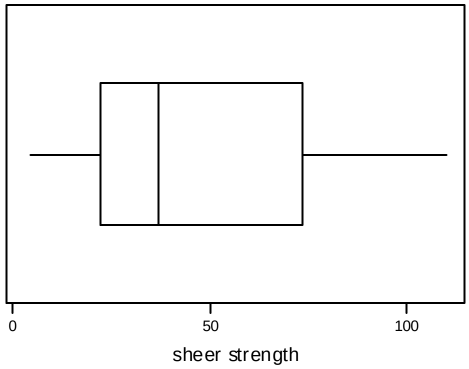
a. The lower half of the data set: 4.4 16.4 22.2 30.0 33.1 36.6, whose median, and therefore, the lower quartile, is $\frac{(22.2+30.0)}{2}+26.1.$

The top half of the data set: 36.6 40.4 66.7 73.7 81.5 109.9, whose median, and therefore, the upper quartile, is $\frac{(66.7+73.7)}{2}=70.2$.

So, the IQR = (70.2 – 26.1) = 44.1

b.

A boxplot (created in Minitab) of this data appears below:



There is a slight positive skew to the data. The variation seems quite large. There are no outliers.

c. An observation would need to be further than 1.5(44.1) = 66.15 units below the lower quartile $[(26.1-66.15)=-40.05\ units]$ or above the upper quartile $[(70.2+66.15)=136.35\ units]$ to be classified as a mild outlier. Notice that, in this case, an outlier on the lower side would not be possible since the sheer strength variable cannot have a negative value.

An extreme outlier would fall (3)44.1) = 132.3 or more units below the lower, or above the upper quartile. Since the minimum and maximum observations in the data are 4.4 and 109.9 respectively, we conclude that there are no outliers, of either type, in this data set.

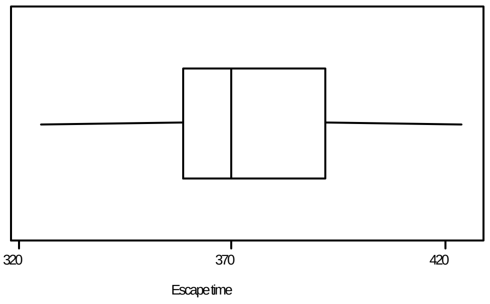
d. Not until the value x = 109.9 is lowered below 73.7 would there be any change in the value of the upper quartile. That is, the value x = 109.9 could not be decreased by more than (109.9 – 73.7) = 36.2 units.

55.

a. Lower half of the data set: 325 325 334 339 356 356 359 359 363 364 364 366 369, whose median, and therefore the lower quartile, is 359 (the 7th observation in the sorted list).
The top half of the data is 370 373 373 374 375 389 392 393 394 397 402 403 424, whose median, and therefore the upper quartile is 392. So, the IQR = 392 - 359 = 33.

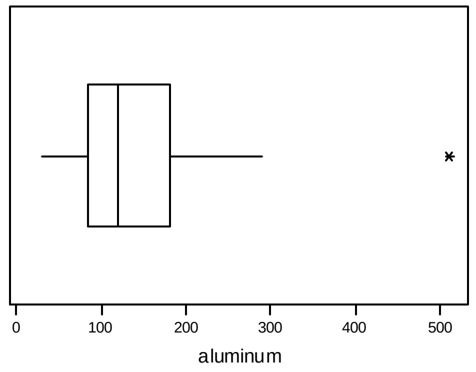
b. 1.5(IQR) = 1.5(33) = 49.5 and 3(IQR) = 3(33) = 99. Observations that are further than 49.5 below the lower quartile (i.e., 359-49.5 = 309.5 or less) or more than 49.5 units above the upper quartile (greater than 392+49.5 = 441.5) are classified as 'mild' outliers. 'Extreme' outliers would fall 99 or more units below the lower, or above the upper, quartile. Since the minimum and maximum observations in the data are 325 and 424, we conclude that there are no mild outliers in this data (and therefore, no 'extreme' outliers either).

c. A boxplot (created by Minitab) of this data appears below. There is a slight positive skew to the data, but it is not far from being symmetric. The variation, however, seems large (the spread 424-325 = 99 is a large percentage of the median/typical value)



d. Not until the value x = 424 is lowered below the upper quartile value of 392 would there be any change in the value of the upper quartile. That is, the value x = 424 could not be decreased by more than 424-392 = 32 units.

56. A boxplot (created in Minitab) of this data appears below.

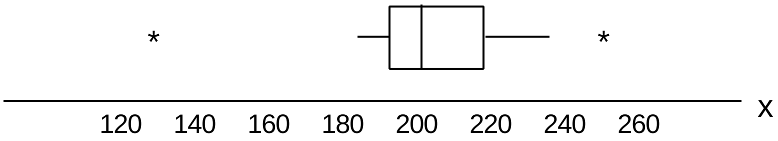


There is a slight positive skew to this data. There is one extreme outler (x=511). Even when removing the outlier, the variation is still moderately large.

57.

a. 1.5(IQR) = 1.5(216.8-196.0) = 31.2 and 3(IQR) = 3(216.8-196.0) = 62.4.
Mild outliers: observations below 196-31.2 = 164.6 or above 216.8+31.2 = 248.
Extreme outliers: observations below 196-62.4 = 133.6 or above 216.8+62.4 = 279.2. Of the observations given, 125.8 is an extreme outlier and 250.2 is a mild outlier.

b. A boxplot of this data appears below. There is a bit of positive skew to the data but, except for the two outliers identified in part (a), the variation in the data is relatively small.



58. The most noticeable feature of the comparative boxplots is that machine 2's sample values have considerably more variation than does machine 1's sample values. However, a typical value, as measured by the median, seems to be about the same for the two machines. The only outlier that exists is from machine 1.

59.

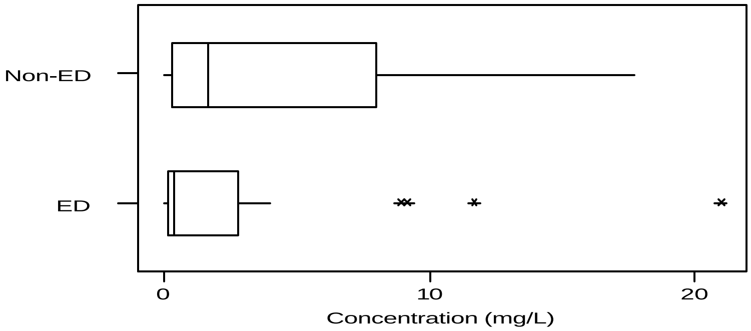
a. ED: median = .4 (the 14th value in the *sorted* list of data). The lower quartile (median of the lower half of the data, including the median, since n is odd) is (.1+.1)/2 = .1. The upper quartile is (2.7+2.8)/2 = 2.75. Therefore, IQR = 2.75 - .1 = 2.65.

Non-ED: median = (1.5+1.7)/2 = 1.6. The lower quartile (median of the lower 25 observations) is .3; the upper quartile (median of the upper half of the data) is 7.9. Therefore, IQR = 7.9 - .3 = 7.6.

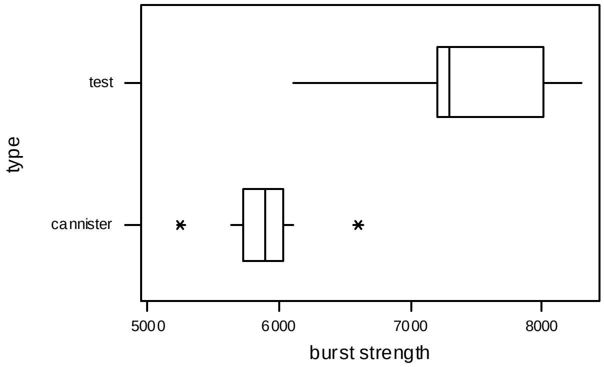
b. ED: mild outliers are less than .1 - 1.5(2.65) = -3.875 or greater than 2.75 + 1.5(2.65) = 6.725. Extreme outliers are less than .1 - 3(2.65) = -7.85 or greater than 2.75 + 3(2.65) = 10.7. So, the two largest observations (11.7, 21.0) are extreme outliers and the next two largest values (8.9, 9.2) are mild outliers. There are no outliers at the lower end of the data.

Non-ED: mild outliers are less than .3 - 1.5(7.6) = -11.1 or greater than 7.9 + 1.5(7.6) = 19.3. Note that there are no mild outliers in the data, hence there can not be any extreme outliers either.

c. A comparative boxplot appears below. The outliers in the ED data are clearly visible. There is noticeable positive skewness in both samples; the Non-Ed data has more variability then the Ed data; the typical values of the ED data tend to be smaller than those for the Non-ED data.



60. A comparative boxplot (created in Minitab) of this data appears below.



The burst strengths for the test nozzle closure welds are quite different from the burst strengths of the production canister nozzle welds.

The test welds have much higher burst strengths and the burst strengths are much more variable.

The production welds have more consistent burst strength and are consistently lower than the test welds. The production welds data does contain 2 outliers.

61. Outliers occur in the 6 a.m. data. The distributions at the other times are fairly symmetric. Variability and the 'typical' values in the data increase a little at the 12 noon and 2 p.m. times.

Supplementary Exercises

62. To somewhat simplify the algebra, begin by subtracting 76,000 from the original data. This transformation will affect each date value and the mean. It will not affect the standard deviation.

$x_1 = 683, \quad x_2 = 1{,}048, \;\; \overline{y} = 831$

$n\overline{x} = (4)(831) = 3{,}324$ so, $x_1 + x_2 + x_3 + x_4 = 3{,}324$

and $x_2 + x_3 = 3{,}324 - x_1 - x_4 = 1{,}593$ and $x_3 = \left(1{,}593 - x_2\right)$

Next, $s^2 = \left(180\right)^2 = \left[\dfrac{\sum x_i^2 - \dfrac{(3324)^2}{4}}{3}\right]$

So, $\sum x_i^2 = 2{,}859{,}444$, $x_1^2 + x_2^2 + x_3^2 + x_4^2 = 2{,}859{,}444$ and

$x_2^2 + x_3^2 = 2{,}859{,}444 - x_1^2 + x_4^2 = 1{,}294{,}651$

By substituting $x_3 = \left(1593 - x_2\right)$ we obtain the equation

$x_2^2 + \left(1{,}593 - x_2\right)^2 - 1{,}294{,}651 = 0$.

$x_x^2 - 1{,}593x_2 + 621{,}499 = 0$

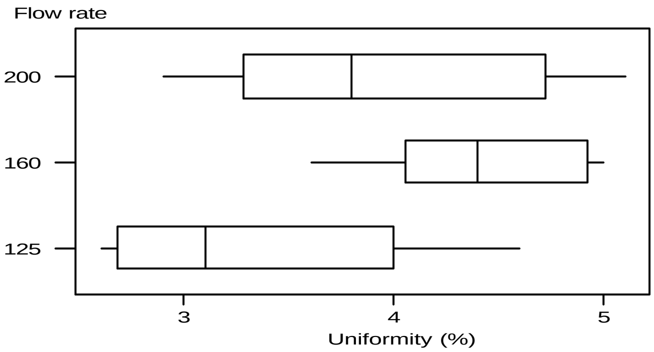
Evaluating for x_2 we obtain $x_2 = 682.8635$ and $x_3 = 1{,}593 - 682.8635 = 910.1365$.

Thus, $x_2 = 76{,}683 \;\; x_3 = 76{,}910$.

63.

Flow rate	**Median**	**Lower quartile**	**Upper quartile**	**IQR**	**1.5(IQR)**	**3(IQR)**
125	3.1	2.7	3.8	1.1	1.65	.3
160	4.4	4.2	4.9	.7	1.05	.1
200	3.8	3.4	4.6	1.2	1.80	3.6

There are no outliers in the three data sets. However, as the comparative boxplot below shows, the three data sets differ with respect to their central values (the medians are different) and the data for flow rate 160 is somewhat less variable than the other data sets. Flow rates 125 and 200 also exhibit a small degree of positive skewness.



64.

6	34
7	17
8	4589
9	1
10	12667789
11	122499
12	2
13	1

stem=ones
leaf=tenths

$\bar{x} = 9.9556, \tilde{x} = 10.6$

$s = 1.7594$

$n = 27$

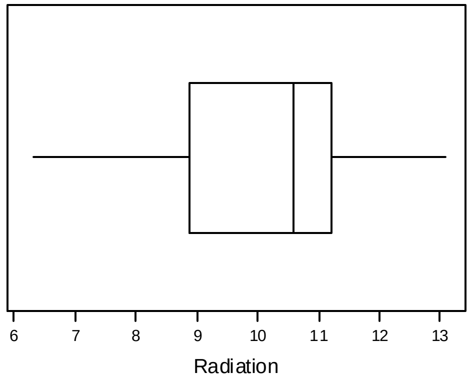
$f_s = 2.3$

lower fourth = 8.85, upper fourth = 11.15

$8.85 - (1.5)(2.3) = 5.4$

$11.15 + (1.5)(2.3) = 14.6$

no outliers



There are no outliers. The distribution is skewed to the left.

65.

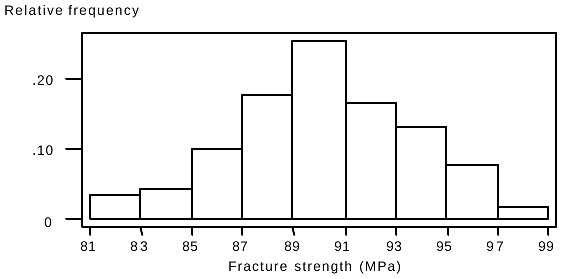
a. HC data: $\sum_i x_i^2 = 2618.42$ and $\sum_i x_i = 96.8$,
so $s^2 = [2618.42 - (96.8)^2/4]/3 = 91.953$
and the sample standard deviation is $s = 9.59$.

CO data: $\sum_i x_i^2 = 145645$ and $\sum_i x_i = 735$, so $s^2 = [145645 - (735)^2/4]/3 = 3529.583$ and the sample standard deviation is $s = 59.41$.

b. The mean of the HC data is 96.8/4 = 24.2; the mean of the CO data is 735/4 = 183.75. Therefore, the coefficient of variation of the HC data is 9.59/24.2 = .3963, or 39.63%. The coefficient of variation of the CO data is 59.41/183.75 = .3233, or 32.33%. Thus, even though the CO data has a larger standard deviation than does the HC data, it actually exhibits *less* variability (in percentage terms) around its average than does the HC data.

66.

a. The histogram appears below. A representative value for this data would be x = 90. The histogram is reasonably symmetric, unimodal, and somewhat bell-shaped. The variation in the data is not small since the spread of the data (99-81 = 18) constitutes about 20% of the typical value of 90.



b. The proportion of the observations that are at least 85 is 1 - (6+7)/169 = .9231. The proportion less than 95 is 1 - (22+13+3)/169 = .7751.

c. x = 90 is the midpoint of the class 89-<91, which contains 43 observations (a relative frequency of 43/169 = .2544. Therefore about half of this frequency, .1272, should be added to the relative frequencies for the classes to the left of x = 90. That is, the approximate proportion of observations that are less than 90 is .0355 + .0414 + .1006 + .1775 + .1272 = .4822.

67.

$$\sum x_i = 163.2$$

$$100\left(\frac{1}{15}\right)\% trimmedmean = \frac{163.2-8.5-15.6}{13} = 10.70$$

$$100\left(\frac{2}{15}\right)\% trimmedmean = \frac{163.2-8.5-8.8-15.6-13.7}{11} = 10.60$$

$$\therefore \frac{1}{2}(100)\left(\frac{1}{15}\right)+\frac{1}{2}(100)\left(\frac{2}{15}\right)=100\left(\frac{1}{10}\right)=10\% trimmedmean$$

$$=\frac{1}{2}(10.70)+\frac{1}{2}(10.60)=10.65$$

68.

a.

$$\frac{d}{dc\left\{\sum (x_i-c)^2\right\}}=\frac{\sum d}{dc(x_i-c)^2}=-2\sum (x_i-c)=0 \Rightarrow \sum (x_i-c)=0$$

$$\Rightarrow \sum x_i - \sum c = 0 \Rightarrow \sum x_i - nc = 0 \Rightarrow nc = \sum x_i \Rightarrow c = \frac{\sum x_i}{n} = \bar{x}.$$

b. $\sum (x_i-\bar{x})^2 \mathit{issmallerthan} \sum (x_i-m)^2.$

69.

a.

$$\bar{y}=\frac{\sum y_i}{n}=\frac{\sum (ax_i+b)}{n}=\frac{a\sum x_i+b}{n}=a\bar{x}+b.$$

$$s_y^2=\frac{\sum (y_i-\bar{y})^2}{n-1}=\frac{\sum (ax_i+b-(a\bar{x}+b))^2}{n-1}=\frac{\sum (ax_i-a\bar{x})^2}{n-1}$$

$$=\frac{a^2\sum (x_i-\bar{x})^2}{n-1}=a^2 s_x^2.$$

b.

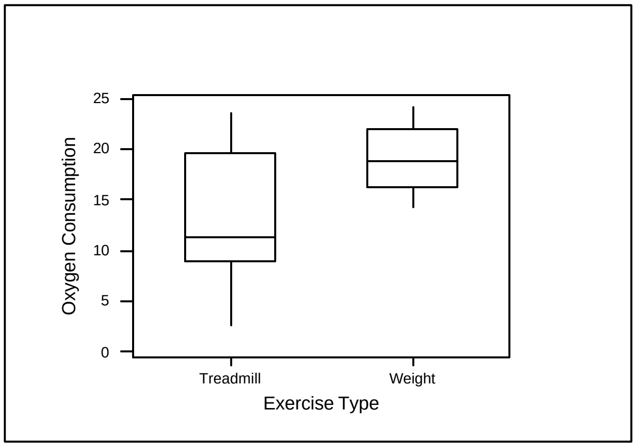
$$x={}^{o}C, y={}^{o}F$$

$$\bar{y}=\frac{9}{5}(87.3)+32=189.14$$

$$s_y=\sqrt{s_y^2}=\sqrt{\left(\frac{9}{5}\right)^2(1.04)^2}=\sqrt{3.5044}=1.872$$

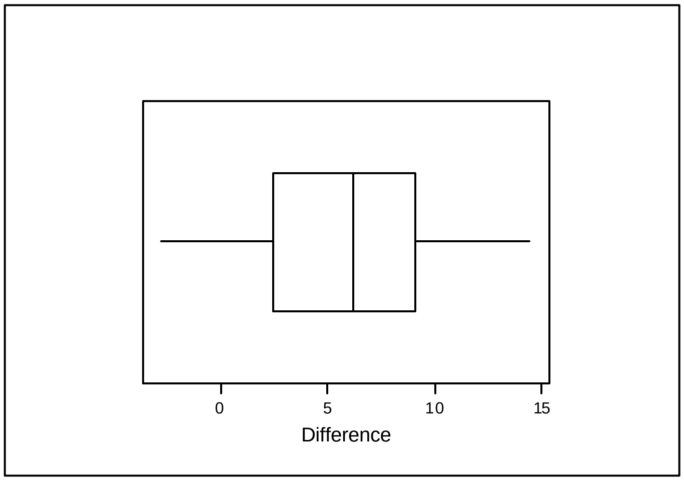
70.

a.



There is a significant difference in the variability of the two samples. The weight training produced much higher oxygen consumption, on average, than the treadmill exercise, with the median consumptions being approximately 20 and 11 liters, respectively.

b. Subtracting the y from the x for each subject, the differences are 3.3, 9.1, 10.4, 9.1, 6.2, 2.5, 2.2, 8.4, 8.7, 14.4, 2.5, -2.8, -0.4, 5.0, and 11.5.

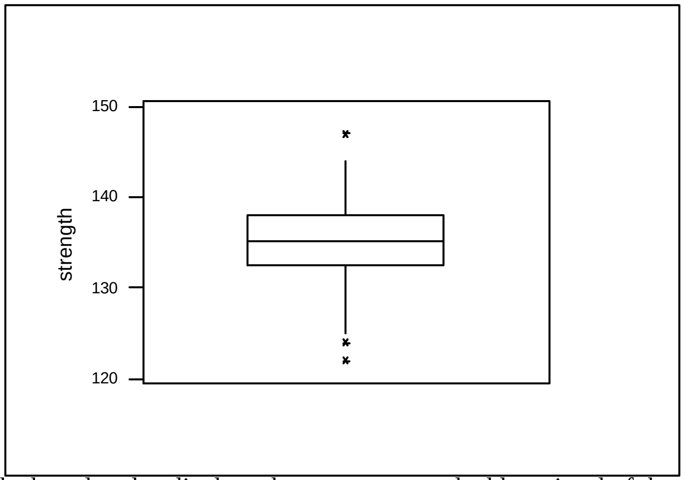


The majority of the differences are positive, which suggests that the weight training produced higher oxygen consumption for most subjects. The median difference is about 6 liters.

71.

a. The mean, median, and trimmed mean are virtually identical, which suggests symmetry. If there are outliers, they are balanced. The range of values is only 25.5, but half of the values are between 132.95 and 138.25.

b.

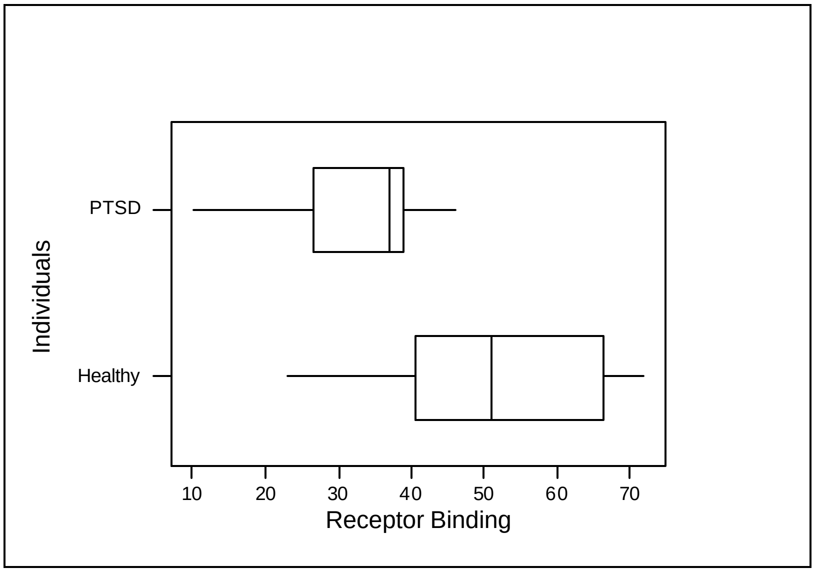


The boxplot also displays the symmetry, and adds a visual of the outliers, two on the lower end, and one on the upper.

72. A table of summary statistics, a stem and leaf display, and a comparative boxplot are below. The healthy individuals have higher receptor binding measure on average than the individuals with PTSD. There is also more variation in the healthy individuals' values. The distribution of values for the healthy is reasonably symmetric, while the distribution for the PTSD individuals is negatively skewed. The box plot indicates that there are no outliers, and confirms the above comments regarding symmetry and skewness.

	PTSD	Healthy
Mean	32.92	52.23
Median	37	51
Std Dev	9.93	14.86
Min	10	23
Max	46	72

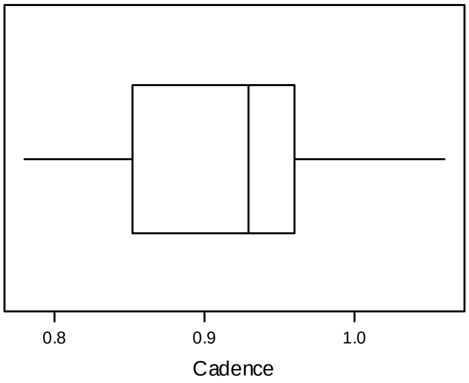
	1	0	stem = tens
3	2	058	leaf = ones
9	3	1578899	
7310	4	26	
81	5		
9763	6		
2	7		



73.

0.7	8	stem=tenths
0.8	11556	leaf=hundredths
0.9	2233335566	
1.0	0566	

$$\overline{x} = .9255, s = .0809, \tilde{x} = .93$$
$$lowerfourth = .855, upperfourth = .96$$



The data appears to be a bit skewed toward smaller values (negatively skewed). There are no outliers. The mean and the median are close in value.

74.

a. Mode = .93. It occurs four times in the data set.

b. The Modal Category is the one in which the most observations occur.

75.

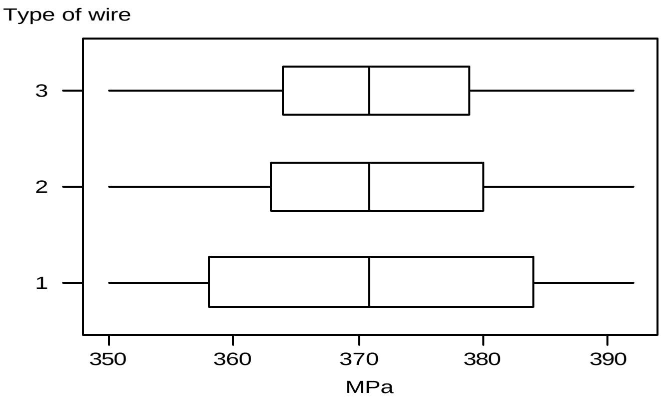
a. The median is the same (371) in each plot and all three data sets are very symmetric. In addition, all three have the same minimum value (350) and same maximum value (392). Moreover, all three data sets have the same lower (364) and upper quartiles (378). So, all three boxplots will be *identical*.

b. A comparative dotplot is shown below. These graphs show that there are differences in the variability of the three data sets. They also show differences in the way the values are distributed in the three data sets.

```
  .                .
  :      .         :::         .      :.
-----+---------+---------+---------+---------+---------+- Type 1

  .  .    .  . .  .. . . . .   .  .   .   .
-----+---------+---------+---------+---------+---------+- Type 2
                         .
  .        . . :. . .    : .:         .
-----+---------+---------+---------+---------+---------+- Type 3
   352.0     360.0     368.0     376.0     384.0     392.0
```

c. The boxplot in (a) is not capable of detecting the differences among the data sets. The primary reason is that boxplots give up some detail in describing data because they use only 5 summary numbers for comparing data sets. Note: The definition of lower and upper quartile used in this text is slightly different than the one used by some other authors (and software packages). Technically speaking, the median of the lower half of the data is not really the first quartile, although it is generally *very close*. Instead, the medians of the lower and upper halves of the data are often called the **lower** and **upper hinges.** Our boxplots use the lower and upper hinges to define the spread of the middle 50% of the data, but other authors sometimes use the *actual* quartiles for this purpose. The difference is usually very slight, usually unnoticeable, but not always. For example in the data sets of this exercise, a comparative boxplot based on the actual quartiles (as computed by Minitab) is shown below. The graph shows substantially the same type of information as those described in (a) except the graphs based on quartiles are able to detect the slight differences in variation between the three data sets.



76. The measures that are sensitive to outliers are: the mean and the midrange. The mean is sensitive because all values are used in computing it. The midrange is sensitive because it uses only the most extreme values in its computation.

The median, the trimmed mean, and the midhinge are not sensitive to outliers.

The median is the most resistant to outliers because it uses only the middle value (or values) in its computation.

The trimmed mean is somewhat resistant to outliers. The larger the trimming percentage, the more resistant the trimmed mean becomes.

The midhinge, which uses the quartiles, is reasonably resistant to outliers because both quartiles are resistant to outliers.

77.

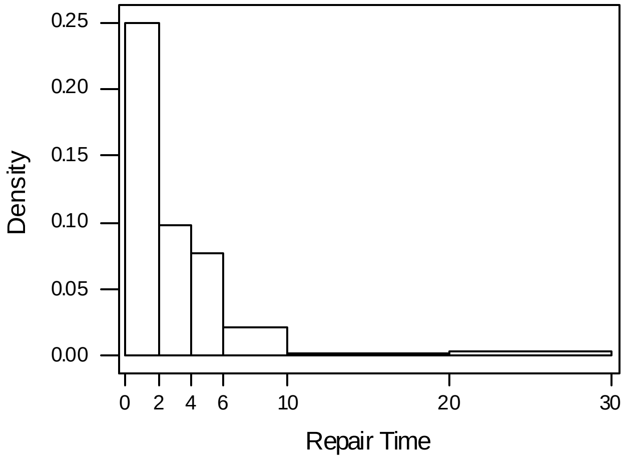
a.

0	2355566777888
1	0000135555
2	00257
3	0033
4	0057
5	044
6	
7	05
8	8
9	0
10	3
HI	22.0 24.5

stem: ones
leaf: tenths

b.

Interval	Frequency	Rel. Freq.	Density
0 -< 2	23	.500	.250
2 -< 4	9	.196	.098
4 -< 6	7	.152	.076
6 -< 10	4	.087	.022
10 -< 20	1	.022	.002
20 -< 30	2	.043	.004



78.

a. Since the constant $\overline{x}$ is subtracted from each x value to obtain each y value, and addition or subtraction of a constant doesn't affect variability, $s_y^2 = s_x^2$ and $s_y = s_x$

b. Let c = 1/s, where s is the sample standard deviation of the x's and also (by a) of the y's. Then $s_z = cs_y = (1/s)s = 1$, and $s_z^2 = 1$. That is, the "standardized" quantities $z_1, \dots, z_n$ have a sample variance and standard deviation of 1.

79.

a. $\sum_{i=1}^{n+1} x_i = \sum_{i=1}^{n} x_i + x_{n+1} = n\overline{x}_n + x_{n+1}, so \overline{x}_{n+1} = \frac{[n\overline{x}_n + x_{n+1}]}{(n+1)}$

b.

$$ns_{n+1}^2 = \sum_{i=1}^{n+1} (x_i - \overline{x}_{n+1})^2 = \sum_{i=1}^{n+1} x_i^2 - (n+1)\overline{x}_{n+1}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2 + x_{n+1}^2 + n\overline{x}_n^2 - (n+1)\overline{x}_{n+1}^2$$

$$= (n-1)s_n^2 + \left\{x_{n+1}^2 + n\overline{x}_n^2 - (n+1)\overline{x}_{n+1}^2\right\}$$

When the expression for $\overline{x}_{n+1}$ from **a** is substituted, the expression in braces simplifies to the following, as desired: $\frac{n(x_{n+1} - \overline{x}_n)^2}{(n+1)}$

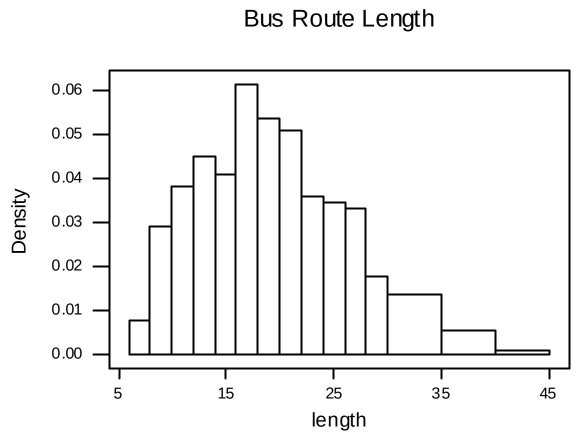
c. $\overline{x}_{n+1} = \frac{15(12.58) + 11.8}{16} = \frac{200.5}{16} = 12.53$

$$s_{n+1}^2 = \frac{n-1}{n}\left(s_n^2\right) + \frac{(x_{n+1} - \overline{x}_n)^2}{(n+1)} = \frac{14}{15}\left(.512^2\right) + \frac{(11.8 - 12.58)^2}{(16)}$$

$= .245 + .038 = .238$. So the standard deviation $s_{n+1} = \sqrt{.238} = .532$

80.

a.



b. Proportion less than $20 = \left(\dfrac{216}{391}\right) = .552$

Proportion at least $30 = \left(\dfrac{40}{391}\right) = .102$

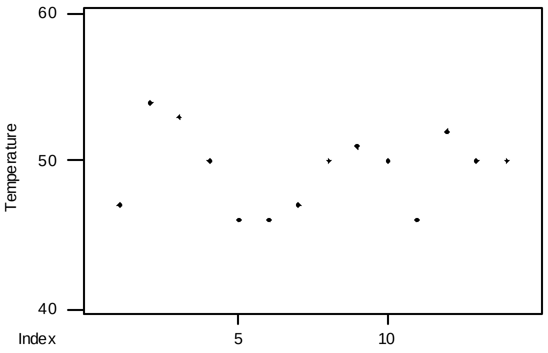
c. First compute (.90)(391 + 1) = 352.8. Thus, the 90th percentile should be about the 352nd ordered value. The 351st ordered value lies in the interval 28 - < 30. The 352nd ordered value lies in the interval 30 - < 35. There are 27 values in the interval 30 - < 35. We do not know how these values are distributed, however, the smallest value (i.e., the 352nd value in the data set) cannot be smaller than 30. So, the 90th percentile is roughly 30.

d. First compute (.50)(391 + 1) = 196. Thus the median (50th percentile) should be the 196 ordered value. The 174th ordered value lies in the interval 16 -< 18. The next 42 observation lie in the interval 18 - < 20. So, ordered observation 175 to 216 lie in the intervals 18 - < 20. The 196th observation is about in the middle of these. Thus, we would say, the median is roughly 19.

81. Assuming that the histogram is unimodal, then there is evidence of positive skewness in the data since the median lies to the left of the mean (for a symmetric distribution, the mean and median would coincide). For more evidence of skewness, compare the distances of the 5th and 95th percentiles from the median: median - 5th percentile = 500 - 400 = 100 while 95th percentile -median = 720 - 500 = 220. Thus, the largest 5% of the values (above the 95th percentile) are further from the median than are the lowest 5%. The same skewness is evident when comparing the 10th and 90th percentiles to the median: median - 10th percentile = 500 - 430 = 70 while 90th percentile -median = 640 - 500 = 140. Finally, note that the largest value (925) is much further from the median (925-500 = 425) than is the smallest value (500 - 220 = 280), again an indication of positive skewness.

82.

a. There is some evidence of a cyclical pattern.



b.

$$\bar{x}_2 = .1x_2 + .9\bar{x}_1 = (.1)(54) + (.9)(47) = 47.7$$

$$\bar{x}_3 = .1x_3 + .9\bar{x}_2 = (.1)(53) + (.9)(47.7) = 48.23 \approx 48.2, etc.$$

t	$\bar{x}_t\ for.a = .1$	$\bar{x}_t\ for.a = .5$
1	47.0	47.0
2	47.7	50.5
3	48.2	51.8
4	48.4	50.9
5	48.2	48.4
6	48.0	47.2
7	47.9	47.1
8	48.1	48.6
9	48.4	49.8
10	48.5	49.9
11	48.3	47.9
12	48.6	50.0
13	48.8	50.0
14	48.9	50.0

α= .1 gives a smoother series.

c.

$$\begin{aligned}\bar{x}_t &= ax_t + (1-a)\bar{x}_{t-1} \\ &= ax_t + (1-a)[ax_{t-1} + (1-a)\bar{x}_{t-2}] \\ &= ax_t + a(1-a)x_{t-1} + (1-a)^2[ax_{t-2} + (1-a)\bar{x}_{t-3}] \\ &= ... = ax_t + a(1-a)x_{t-1} + a(1-a)^2 x_{t-2} + ... + a(1-a)^{t-2}x_2 + (1-a)^{t-1}\bar{x}_1\end{aligned}$$

Thus, (x bar)$_t$ depends on x_t and all previous values. As k increases, the coefficient on x_{t-k} decreases (further back in time implies less weight).

d. Not very sensitive, since $(1-\alpha)^{t-1}$ will be very small.

83.

a. When there is perfect symmetry, the smallest observation y_1 and the largest observation y_n will be equidistant from the median, so $y_n - \overline{x} = \overline{x} - y_1$.

Similarly, the second smallest and second largest will be equidistant from the median, so $y_{n-1} - \overline{x} = \overline{x} - y_2$

and so on. Thus, the first and second numbers in each pair will be equal, so that each point in the plot will fall exactly on the 45 degree line. When the data is positively skewed, y_n will be much further from the median than is y_1, so $y_n - \tilde{x}$ will considerably exceed $\tilde{x} - y_1$ and the point $(y_n - \tilde{x}, \tilde{x} - y_1)$ will fall considerably below the 45 degree line. A similar comment aplies to other points in the plot.

b. The first point in the plot is (2745.6 – 221.6, 221.6 0- 4.1) = (2524.0, 217.5). The others are: (1476.2, 213.9), (1434.4, 204.1), (756.4, 190.2), (481.8, 188.9), (267.5, 181.0), (208.4, 129.2), (112.5, 106.3), (81.2, 103.3), (53.1, 102.6), (53.1, 92.0), (33.4, 23.0), and (20.9, 20.9). The first number in each of the first seven pairs greatly exceed the second number, so each point falls well below the 45 degree line. A substantial positive skew (stretched upper tail) is indicated.