Solutions Manual for Probability and Statistics for Engineering and the Sciences 10th Edition by Ellingson, Panorska

ISBN: 9798214023823

Solution and Answer Guide

ELLINGSON AND PANORSKA, PROBABILITY AND STATISTICS FOR ENGINEERING AND THE SCIENCES, 10E, 2027, 9798214023823; CHAPTER 1: OVERVIEW AND DESCRIPTIVE STATISTICS

TABLE OF CONTENTS

SECTION 1.1

1.

Solutions:

a. Los Angeles Times, Chicago Tribune, Gainesville Sun, Washington Post

b. France, Sweden, Denmark, Malta

c. Alice Johnson, Catherine Miller, Emma Davis, Ken Lee

d. 2.97, 3.56, 2.20, 2.97

2.

Solutions:

a. 45.3 m, 52.7 m, 60.1 m, 48.9 m

b. 432 pp, 196 pp, 184 pp, 321 pp

c. 4.5, 4.9, 5.6, 6.4

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

3.

Solutions:

a. How likely is it that more than half of the sampled computers will need or have needed warranty service? What is the expected number among the 100 that need warranty service? How likely is it that the number needing warranty service will exceed the expected number by more than 10?

b. Suppose that 15 of the 100 sampled needed warranty service. How confident can we be that the proportion of *all* such computers needing warranty service is between .08 and .22? Does the sample provide compelling evidence for concluding that more than 10% of all such computers need warranty service?

4.

Solutions:

a. Concrete populations: all living citizens of U.S., all the mutual funds currently marketed in U.S., all the books that were published in 1980. Hypothetical populations: All GPAs of the University of California undergraduates in the next academic year, page lengths of all books that will be published next year, the batting averages of all major league players in 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%? (Hypothetical) 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.

Solutions:

a. No. All students taking a large statistics course who participate in an SI program of this sort.

b. The advantage to randomly allocating students to the two groups is that the two groups should then be fairly comparable before the study. If the two groups perform differently in the class, we might attribute this to the treatments (SI and control). If students were allowed to choose, stronger or more motivated students might prefer SI, which could bias the results.

c. If all students were put in the treatment group, there would be no firm basis for assessing the effectiveness of SI (nothing to which the SI scores could reasonably be compared).

6.

Solution:

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 from 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.

Solution:

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 selected homes, values of all desired variables would be

determined. This is an enumerative study as there is a finite, well-defined population of objects that can be used as a sample.

8.

Solutions:

a. Number observations equal $2 \times 2 \times 2 = 8$

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

9.

Solutions:

a. The measurements might vary for different reasons. Some of these include measurement error resulting from mechanical or technical changes across measurements, recording errors, differences in weather conditions at the time of measurements, etc.

b. No, because there is no sampling frame.

SECTION 1.2

10.

Solutions:

a.

5	9
6	33588
7	00234677889
8	127
9	077
10	7
11	368

stem: ones
leaf: tenths

A representative strength for these beams is around 7.8 MPa, but there is a reasonably large amount of variation around that representative value. (What constitutes large or small variation usually depends on context, but variation is usually considered large when the range of the data – the difference between the largest and smallest value – is comparable to a representative value. Here, the range is $11.8 - 5.9 = 5.9$ MPa, which is similar in size to the representative value of 7.8 MPa. So, most researchers would call this a large amount of variation.)

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

c. 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. (A later section 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.

Solution:

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

The stem-and-leaf plot shows that .45 is a good representative value for the data. It also shows that the plot is not symmetric and it is positively skewed. The range of the data is $.75 - .31 = .44,$ which is comparable to the typical value of .45. This constitutes a reasonably large amount of variation in the data. The data value .75 is a possible outlier.

12.

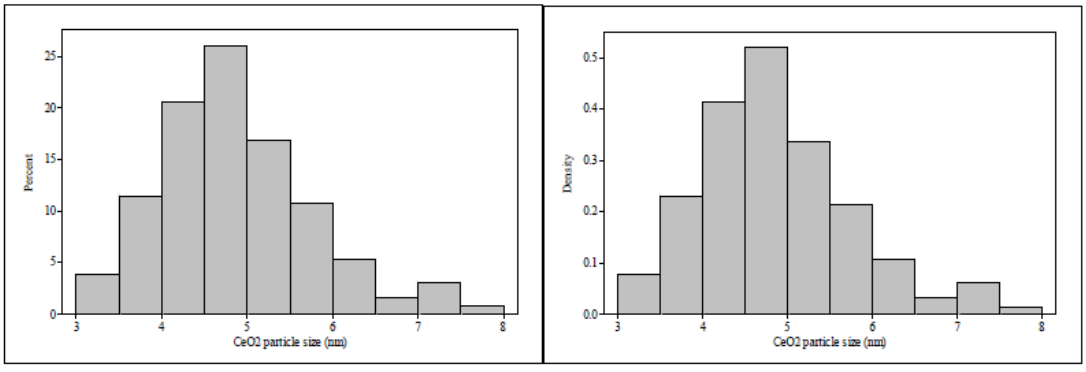
Solutions:

The sample size for this data set is $n = 5 + 15 + 27 + 34 + 22 + 14 + 7 + 2 + 4 + 1 = 131.$

a. The first four intervals correspond to observations less than 5, so the proportion of values less than 5 is $\left(5 + 15 + 27 + 34\right)/131 = 81/131 = .618.$

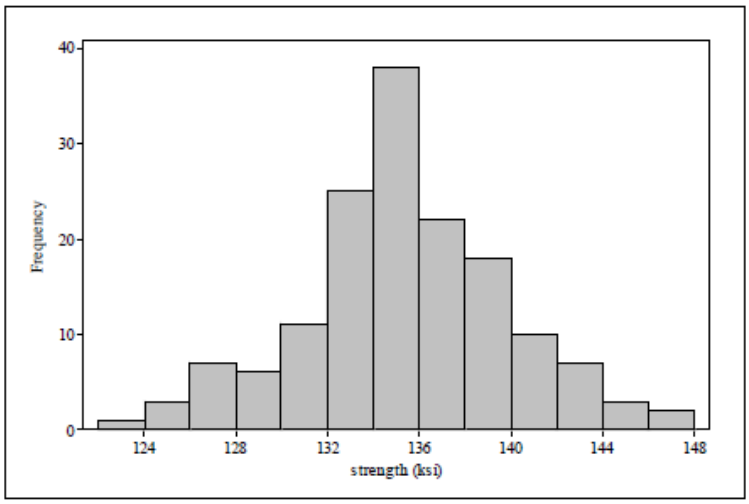
b. The last four intervals correspond to observations at least 6, so the proportion of values at least 6 is $\left(7 + 2 + 4 + 1\right)/131 = 14/131 = .107.$

c. The relative frequency and density histograms are shown below. The distribution of CeO_2 particle sizes is not symmetric, but are positively skewed. Notice that the relative frequency and density histograms are essentially identical, other than the vertical axis labeling, because the bin widths are all the same.



13.

Solution:



The histogram of ultimate strengths is symmetric and unimodal, with the point of symmetry at approximately 135 ksi. There is a moderate amount of variation, and there are no gaps or outliers in the distribution.

14.

Solutions:

a.

2	23	stem: 1.0
3	2344567789	leaf: .10
4	01356889	
5	00001114455666789	
6	000012222334445666778999 9	
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 is around 7.0.

c. The data exhibit a moderate amount of variation (this is subjective).

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.

Solution:

American		French
	8	1
755543211000	9	00234566
9432	10	2356
6630	11	1369
850	12	223558
8	13	7
	14	
	15	8
2	16	

American movie times are unimodal strongly positively skewed, while French movie times are bimodal. A typical American movie runs about 95 minutes, while French movies are typically either around 95 minutes or around 125 minutes. American movies are generally shorter than French movies and are less variable in length. Finally, both American and French movies occasionally run very long (outliers at 162 minutes and 158 minutes, respectively, in the samples).

16.

Solutions:

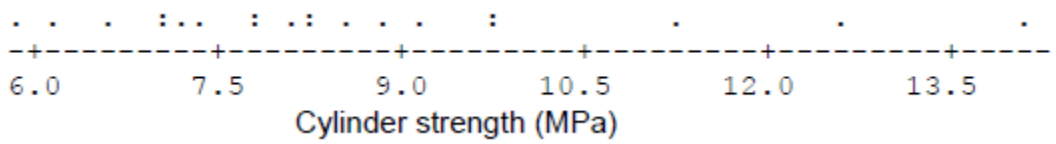
a.

Beams		Cylinders	
9	5	8	
88533	6	16	
98877643200	7	012488	
721	8	13359	stem: ones
770	9	278	leaf: tenths
7	10		
863	11	2	
	12	6	
	13		
	14	1	

The data are slightly skewed to the right, or positively skewed. The value of 14.1 MPa is an outlier. About 15% of the measurements, that is, three out of twenty are 10 MPa.

b. The majority of observations are between 5 and 9 MPa for both beams and cylinders, with the modal class being 7.0-7.9 MPa. The observations for cylinders arespread out, and the maximum value of the cylinder observations is higher.

c.



17.

Solutions:

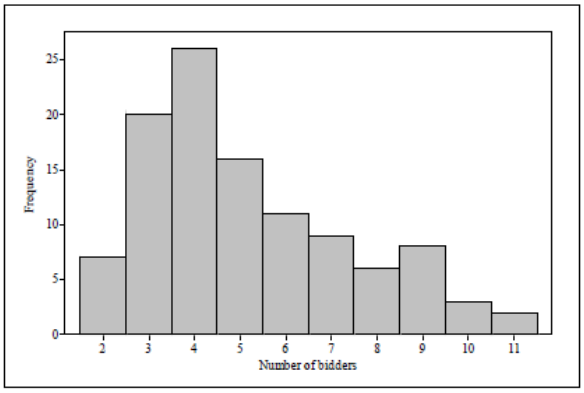
The sample size for this data set is $n = 7 + 20 + 26 + \cdots + 3 + 2 = 108$.

a. "At most five bidders" means 2, 3, 4, or 5 bidders. The proportion of contracts that involved at most 5 bidders is $(7 + 20 + 26 + 16)/108 = 69/108 = .639$.

Similarly, the proportion of contracts that involved at least 5 bidders (5 through 11) is equal to $(16 + 11 + 9 + 6 + 8 + 3 + 2)/108 = 55/108 = .509$.

b. The number of contracts with between 5 and 10 bidders, inclusive, is $16 + 11 + 9 + 6 + 8 + 3 = 53$, so the proportion is $53/108 = .491$. "Strictly" between 5 and 10 means 6, 7, 8, or 9 bidders, for a proportion equal to $(11 + 9 + 6 + 8)/108 = 34/108 = .315$.

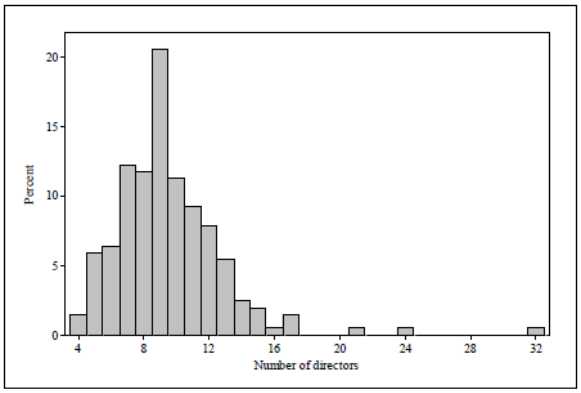
c. The distribution of number of bidders is positively skewed, ranging from 2 to 11 bidders, with a typical value of around 4-5 bidders.



18.

Solutions:

a. The most interesting feature of the histogram is the three very large outliers (21, 24, and 32 directors). Without these, the distribution of number of directors would be roughly symmetric with a typical value of around 9.



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, 42(9) to enter 42 copies of the number 9. The frequency data in this exercise was entered using the following Minitab commands:

```
MTB > set c1
DATA> 3(4) 12(5) 13(6) 25(7) 24(8) 42(9) 23(10) 19(11) 16(12)
11(13) 5(14) 4(15) 1(16) 3(17) 1(21) 1(24) 1(32)
DATA> end
```

b. The frequency distribution shown below is almost the same as the one in the textbook, except the three largest values are grouped into the "≥ 18" category. If the data is presented this way, we could not create a histogram, because we wouldn't know the upper boundary for the rectangle corresponding to the "≥ 18" category.

No. dir.	4	5	6	7	8	9	10	11
Freq.	3	12	13	25	24	42	23	19

No dir.	12	13	14	15	16	17	≥ 18
Freq.	16	11	5	4	1	3	3

c. The sample size is $3+12+\cdots+3+1+1+1=204$. So, the proportion of these corporations that have at most 10 directors is $(3+12+13+25+24+42+23)/204 = 142/204 = .696$.

d. Similarly, the proportion of these corporations with more than 15 directors is $(1+3+1+1+1)/204 = 7/204 = .034$.

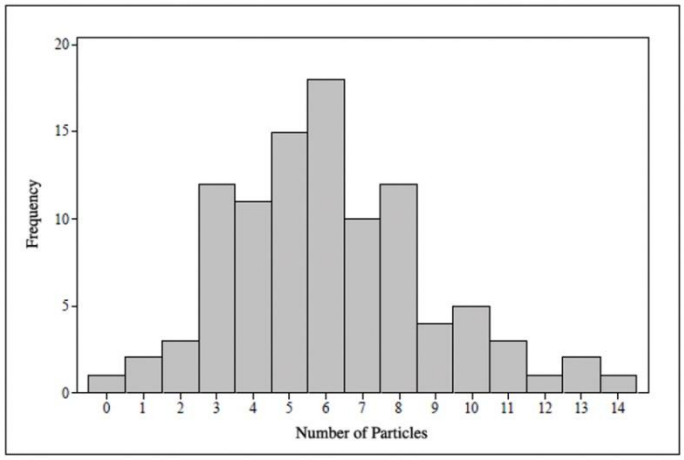
19.

Solutions:

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 $1, 2, 3, \ldots$ particles. In a similar fashion, the proportion containing at least 5 particles is $\left(100-1-2-3-12-11\right)/100 = 71/100 = .71,$ or, 71%.

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

c. The following histogram was constructed using Minitab. It is *almost* symmetric and unimodal. It has a few smaller modes and has a very slight positive skew.



20.

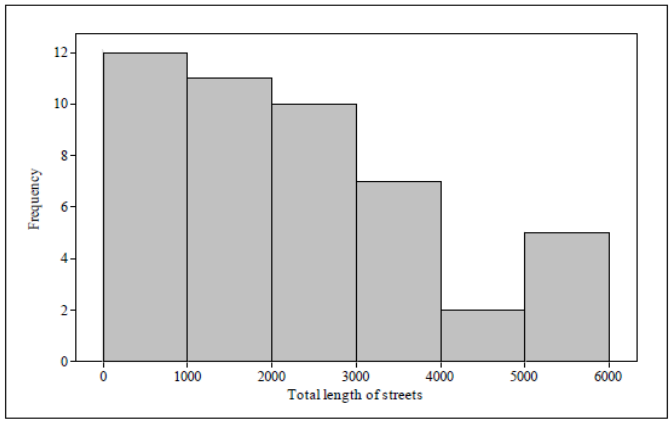
Solutions:

a. The following stem-and-leaf plot 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 bimodal (the stem at 5 would be considered a mode, the stem at 0 another) and has a positive skew.

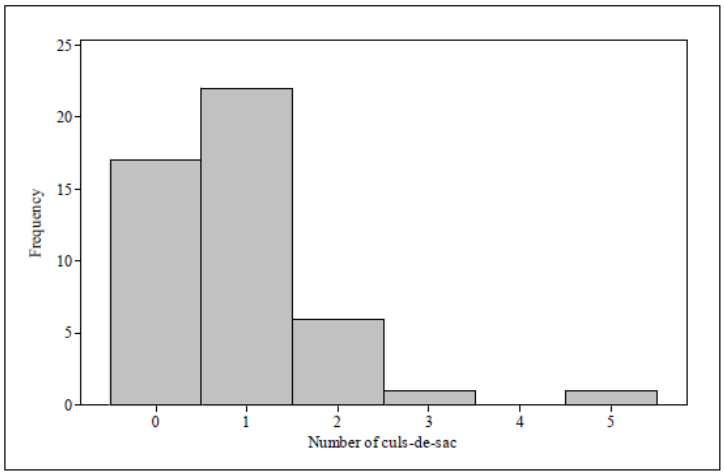
b. A histogram of this data, using classes boundaries of $0, 1000, 2000, \ldots, 6000$ is shown below. The proportion of subdivisions with total length less than 2000 is $(12 + 11)/47 = .489$, or 48.9%. Between 2000 and 4000, the proportion is $(10 + 7)/47 = .362$, or 36.2%. The histogram shows the same general shape as depicted by the stem-and-leaf in part (a).



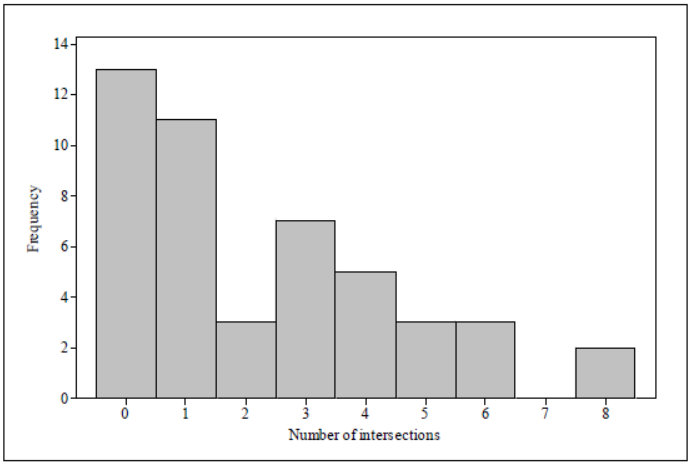
21.

Solutions:

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 $(y \geq 1)$ 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 $y \geq 1$.



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



22.

Solution:

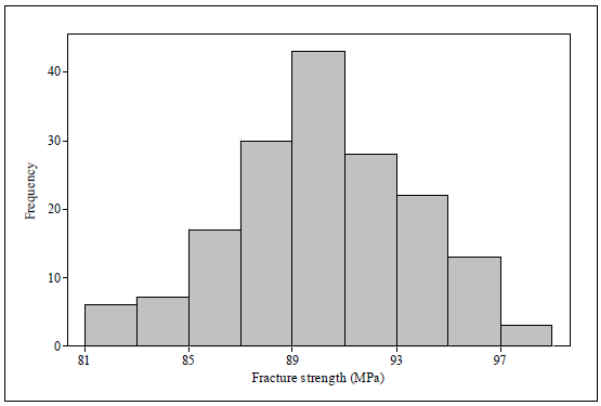
A very large percentage of the data values are greater than 0, which indicates that most, but not all, runners do slow down at the end of the race. The histogram is also positively skewed, which means that some runners slow down a *lot* compared to the

others. A typical value for this data is in the neighborhood of 200 seconds. The proportion of the runners who ran the late distance more quickly than the early distance is very small, about 1% or so.

23.

Solutions:

a. The histogram is shown below. A representative value for this data is around 90 MPa. The histogram is reasonably symmetric, unimodal, and somewhat bell-shaped with a fair amount of variability ($s \approx 3$ or 4 MPa).



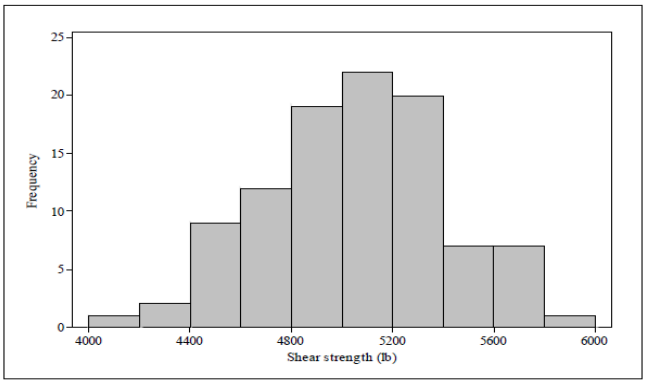
b. The proportion of the observations that are at least 85 is $1-(6+7)/169 = .9231$. The proportion less than 95 is $1-(13+3)/169 = .9053$.

c. 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 $.0355 + .0414 + .1006 + .1775 + .1272 = .4822$.

24.

Solution:

The distribution of shear strengths is roughly symmetric and bell-shaped, centered at about 5000 lbs and ranging from about 4000 to 6000 lbs.

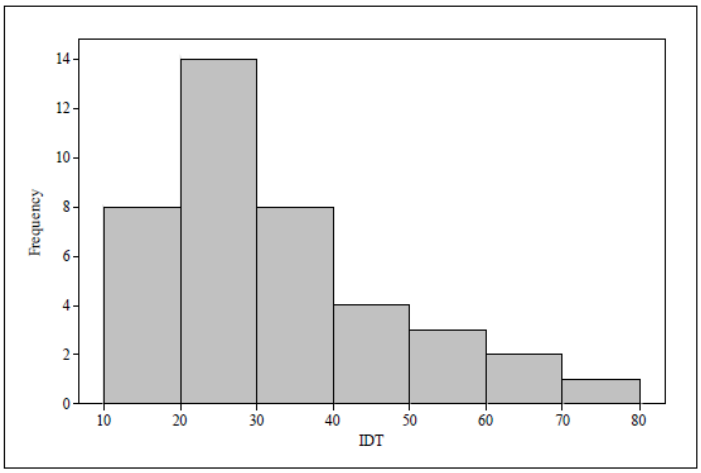


25.

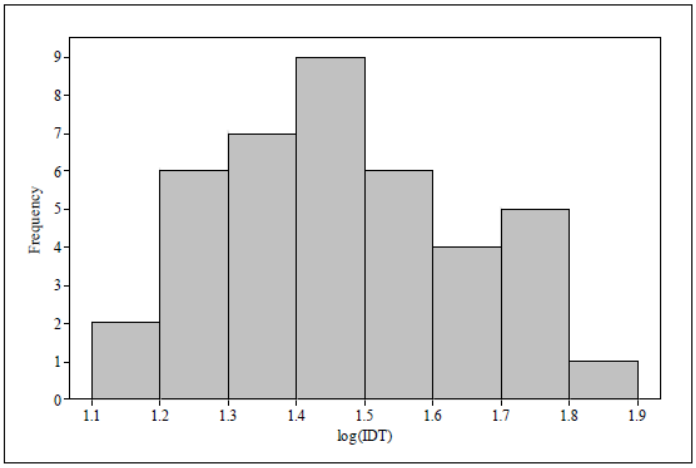
Solution:

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

Histogram of original data:



Histogram of transformed data:

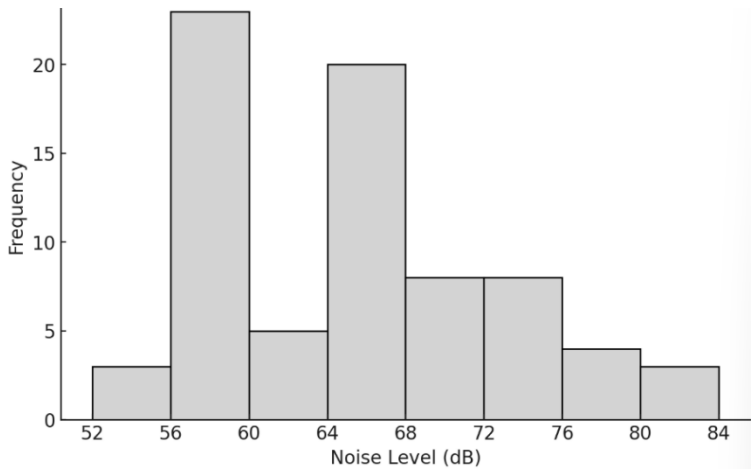


26.

Solutions:

a. Stem | Leaf
5 | 55566666666667777888888999
6 | 022444444555555555777799999
7 | 00111333355559999
8 | 333

b.



c. As seen in the histogram, this noise distribution is bimodal (but close to unimodal) with a positive skew. Most of the observations are in the lower noise level (55-60 dBA). There are also few observations in the higher ranges (64-68 dBA) which contribute to the negative skew.

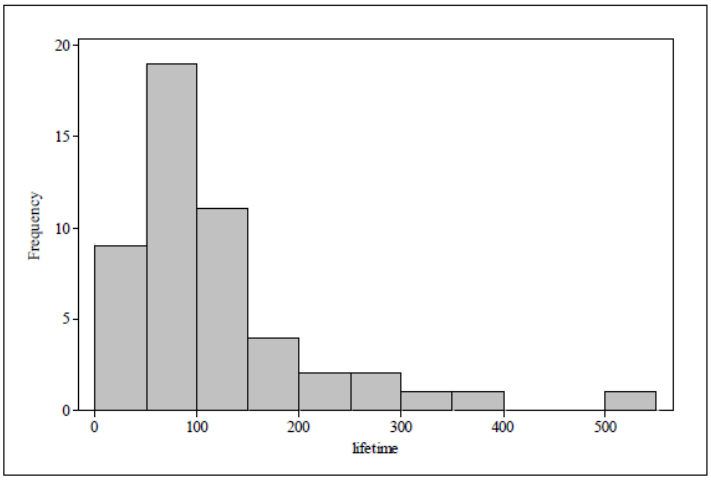
27.

Solutions:

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

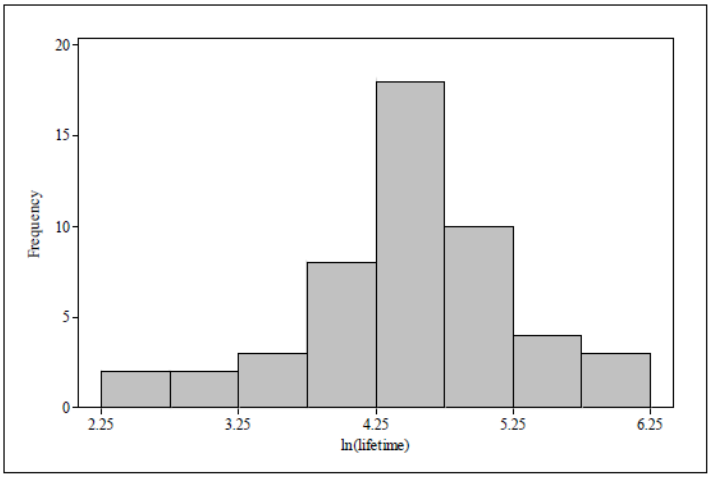
b. The lifetime distribution is positively skewed. A representative value is around 100. There is a great deal of variability in lifetimes and several possible candidates for outliers.

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–<450	0	0.00
450–<500	0	0.00
500–<550	1	0.02
	50	1.00



c. There is much more symmetry in the distribution of the transformed values than in the values themselves, and less variability. There are no longer gaps or obvious outliers.

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



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.

Solutions:

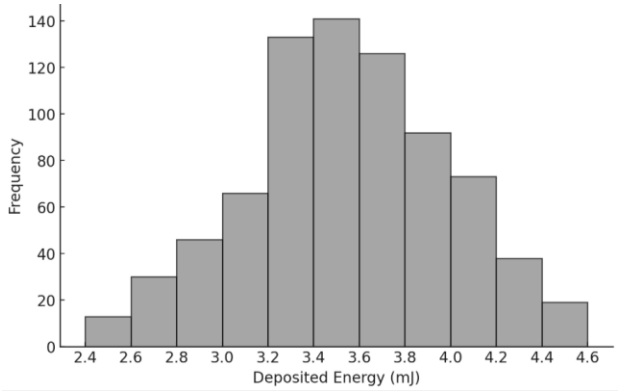
The sample size for this data set is n = 777

a. (13 + 30 + 46)/777 = 0.115

b. (73 + 38 + 19)/777 = 0.167

c. The number of trials resulting in deposited energy of 3.5 mJ or more is (141 + 126 + 92 + 73 + 38 + 19)/777 = 0.629.

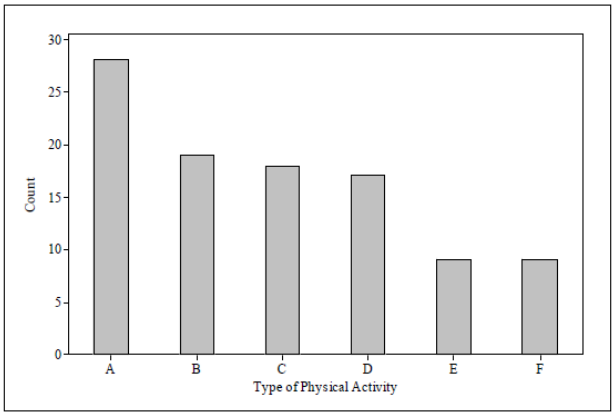
d.



29.

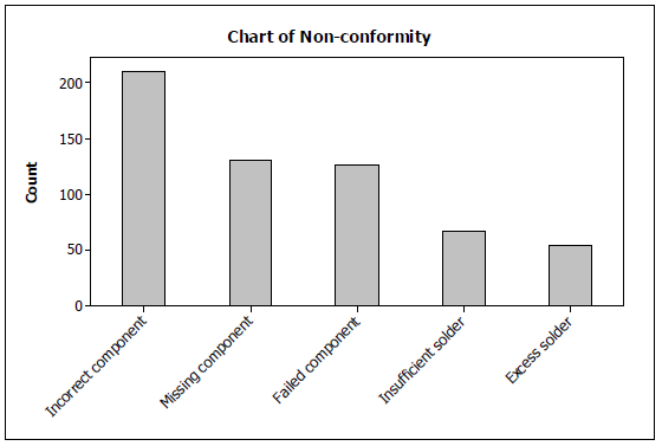
Solution:

Physical Activity	Frequency	Relative Frequency
A	28	.28
B	19	.19
C	18	.18
D	17	.17
E	9	.09
F	9	.09
	100	1.00



30.

Solution:



31.

Solutions:

a.

Class	Freq.	Cum. freq.	Cum. Rel. freq.
3.0-<3.5	5	5	.0382
3.5-<4.0	15	20	.1527
4.0-<4.5	27	47	.3588
4.5-<5.0	34	81	.6183
5.0-<5.5	22	103	.7863
5.5-<6.0	14	117	.8931
6.0-<6.5	7	124	.9466
6.5-<7.0	2	126	.9618
7.0-<7.5	4	130	.9924
7.5-<8.0	1	131	1.0000

b.

No. of Directors	Frequency (f)	Cum. freq.	Cum. Rel. freq.
4	3	2	.0147
5	12	15	.0735
6	13	27	.1373
7	25	53	.2598
8	24	77	.3775
9	42	119	.5833
10	23	142	.6961
11	19	161	.7892
12	16	177	.8676
13	11	188	.9216
14	5	193	.9461
15	4	197	.9657
16	1	198	.9706
17	3	201	.9853

21	1	202	.9902
24	1	203	.9951
32	1	204	1.0000

c.

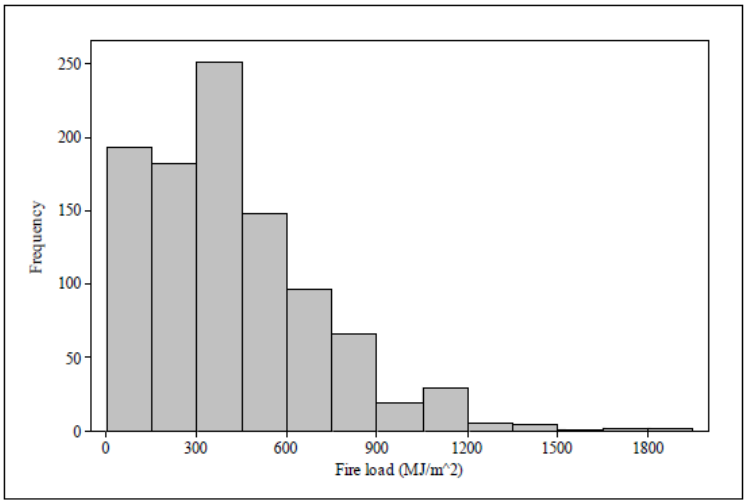
Class Interval	Freq.	Cum. freq.	Cum. Rel. Freq.
4000-4200	1	1	.011
4200-4400	3	4	.044
4400-4600	8	12	.133
4600-4800	12	24	.267
4800-5000	15	39	.433
5000-5200	18	57	.633
5200-5400	14	71	.789
5400-5600	10	81	.900
5600-5800	6	87	.967
5800-6000	3	90	1.000

32.

Solutions:

a. First, the cumulative percents must be converted to relative frequencies. Then the histogram may be constructed (see below). The relative frequency distribution is almost unimodal and shows 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.

Class	Rel. Freq.	Class	Rel. Freq.
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



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.

Solutions:

a. Using software, $\bar{x}=640.5\ (\$640{,}500)$ and. $\tilde{x}=582.5\ (\$582{,}500)$. The average sale price for a home in this sample was \$640,500. Half the sales were for less than \$582,500, while half were for more than \$582,500.

b. Changing that one value lowers the sample mean to 610.5 (\$610,500) but has no effect on the sample median.

c. After removing the two largest and two smallest values, $\bar{x}_{tr(20)}=(\$591{,}200)$.

d. A 10% trimmed mean from removing just the highest and lowest values is $\bar{x}_{tr(10)}=596.3$. To form a 15% trimmed mean, take the average of the 10% and 20% trimmed means to get $\bar{x}_{tr(15)}=(591.2+596.3)/2=593.75\ (\$593{,}750)$.

34.

Solutions:

a. Sample data:

Seker (S): 610.3, 638.0, 624.1, 645.9, 620.1, 634.9, 670.0, 629.7, 686.6, 675.4

Dermason (D): 653.9, 659.0, 647.2, 642.6, 645.8, 651.6, 650.4, 647.9, 649.1, 638.9

The mean of the Seker sample is:

$$\bar{x}_S=\frac{\sum S}{10}$$

$$\bar{x}_S=\frac{610.3+638.0+624.1+645.9+620.1+634.9+670.0+629.7+686.6+675.4}{10}$$

$$\bar{x}_S=\frac{6435.0}{10}$$

$$\bar{x}_S=643.5$$

The mean of the Dermason sample is:

$$\bar{x}_D=\frac{\sum D}{10}$$

$$\bar{x}_D=\frac{653.9+659.0+647.2+642.6+645.8+651.6+650.4+647.9+649.1+638.9}{10}$$

$$\bar{x}_D = \frac{6486.4}{10}$$

$$\bar{x}_D = 648.64$$

Comparison:
The mean perimeter for Dermason beans (648.64) is slightly larger than for Seker beans (643.5). Thus, Dermason beans appear on average a bit longer in outline.

b. Sorted Seker (S) sample values:

610.3, 620.1, 624.1, 629.7, 634.9, 638.0, 645.9, 670.0, 675.4, 686.6

For $n = 10$, the median is the average of the 5th and 6th values:

$$\text{Median}_S = \frac{634.9 + 638.0}{2} = \frac{1272.9}{2} = 636.45$$

Sorted Dermason (D) sample values:
638.9, 642.6, 645.8, 647.2, 647.9, 649.1, 650.4, 651.6, 653.9, 659.0

The median is the average of the 5th and 6th values:

$$\text{Median}_D = \frac{647.9 + 649.1}{2} = \frac{1297.0}{2} = 648.5$$

Comparison:
The median for Dermason beans (648.5) is again larger than the median for Seker beans (636.5). The difference is slightly greater here than in the means.

c. We remove 1 observation at each end (smallest and largest). With 10 values, trimming one from each side removes $\frac{2}{10} = 20\%$ of the data (10% from each side).

Seker (S):

Remove smallest (610.3) and largest (686.6).

Remaining 8 values: 620.1, 624.1, 629.7, 634.9, 638.0, 645.9, 670.0, 675.4

$$\text{Trimmed Mean}_S = \frac{620.1 + 624.1 + 629.7 + 634.9 + 638.0 + 645.9 + 670.0 + 675.4}{8}$$

$$\text{Trimmed Mean}_S = \frac{5138.1}{8} \approx 642.26$$

Dermason (D):

Remove smallest (638.9) and largest (659.0).

Remaining 8 values: 642.6, 645.8, 647.2, 647.9, 649.1, 650.4, 651.6, 653.9

$$\text{Trimmed Mean}_D = \frac{642.6 + 645.8 + 647.2 + 647.9 + 649.1 + 650.4 + 651.6 + 653.9}{8}$$

$$\text{Trimmed Mean}_D = \frac{5188.5}{8} \approx 648.56$$

Comparison:

- Trimmed mean for Seker = 642.26.0, slightly lower than its mean (643.5) and close to its median (636.5).
- Trimmed mean for Dermason = 648.6, nearly the same as both its mean (648.6) and its median (648.5).

This suggests the Seker data are a bit more affected by extreme values (like 610.3 and 686.6), whereas the Dermason data are more symmetric and balanced.

35.

Solutions:

The sample size is $n = 15$.

a. The sample mean is $\bar{x} = 18.55 / 15 = 1.237\ \mu g/g$ and the sample median is $\tilde{x} =$ the 8^{th} ordered value = .56 µg/g. These values are very different due to the heavy positive skewness in the data.

b. The 6.67% trimmed mean calculated by removing the highest and lowest values is 1.162. This trimmed mean is less than mean because it is less spread out and it is larger than the value of median.

c. The median of the data set will remain .56 so long as that's the 8^{th} ordered observation. Hence, the value .20 could be increased to as high as .56 without changing the fact that the 8^{th} ordered observation is .56. Equivalently, .20 could be increased by as much as .36 without affecting the value of the sample median.

d. In 6.67% trimmed mean, only the lowest and highest values are removed. Here 0.20 is the lowest value. So, 0.20 can be increased as long as it is less than the second lowest value 0.22.

36.

Solutions:

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 $\bar{x} = 9638/26 = 370.7$ sec, while the sample median is $\tilde{x} = (369 + 370)/2 = 369.50$ sec.

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 370, and hence, the median will not change. However, the value $x = 424$ cannot be changed to a number less than 370 (a change of $424 - 370 = 54$) since that will change the middle two values.

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

37.

Solutions:

a. $\bar{x} = 12.01$, $\tilde{x} = 11.35$, $\bar{x}_{tr(10)} = 11.46$.

b. The median or the trimmed mean would be better choices than the mean because of the outlier 21.9.

38.

Solutions:

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.

Solutions:

a. $\sum x_i = 16.475$ so $\bar{x} = \dfrac{16.475}{16} = 1.0297$; $\tilde{x} = \dfrac{(1.007 + 1.011)}{2} = 1.009$

b. 1.394 can be decreased until it reaches 1.011 (i.e. by $1.394 - 1.011 = 0.383$), the largest of the 2 middle values. If it is decreased by more than 0.383, the median will change.

40.

Solution:

$\tilde{x} = 60.8$, $\bar{x}_{tr(25)} = 59.3083$, $\bar{x}_{tr(10)} = 58.3475$, $\bar{x} = 58.54$. All four measures of center have about the same value.

41.

Solutions:

a. $\hat{p} = x/n = 7/10 = .7$

b. $\bar{x} = .70 =$ the sample proportion of successes ($\hat{p}$)

c. To have $\hat{p}$ equal .80 requires $x/25 = .80$ or $x = (.80)(25) = 20$. There are 7 successes (S) already, so another $20 - 7 = 13$ would be required.

42.

Solutions:

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

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

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

$\tilde{y}$ = the median of $\left(cx_1, cx_2, \ldots, cx_n\right) = c \cdot$ the median of $(x_1, x_2, \ldots, x_n) = c\tilde{x}$

43.

Solution:

The median and certain trimmed means can be calculated, while the mean cannot — the exact values of the "100+" observations are required to calculate the mean.

$\tilde{x} = \dfrac{(57 + 79)}{2} = 68.0$, $\bar{x}_{tr(20)} = 66.2$, $\bar{x}_{tr(30)} = 67.5$.

44.

Solution:

$$\sum_{i=1}^{n+1} x_i = \sum_{i=1}^{n} x_i + x_{n+1} = n\bar{x}_n + x_{n+1}, \text{ so } \bar{x}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} x_i = \frac{n\bar{x}_n + x_{n+1}}{n+1}.$$

45.

Solutions:

a. The mode(s) of the data are 356, 359, 325, 373, and 364, as they each appear more than once in the sample.

b. There is no mode for the given data.

c. There are two modes 120 and 130 both of which are repeated two times.

46.

Solutions:

a. S is the mode for this data set which is repeated seven times.

b. A is the mode for the data set which is repeated 28 times.

47.

Solutions:

a. From exercise 45, the total contracts are 108 (sum of all contracts).

The mean (average number of bidders per contract) is given by the weighted average:

$$\bar{x} = \frac{\sum(\text{Number of Bidders} \times \text{Number of Contracts})}{\text{Total Number of Contracts}}$$

$$\bar{x} = \frac{(2\times 7)+(3\times 20)+(4\times 26)+(5\times 16)+(6\times 11)+(7\times 9)+(8\times 6)+(9\times 8)+(10\times 3)+(11\times 2)}{108}$$

$$\bar{x} = \frac{14+60+104+80+66+63+48+72+30+22}{108}$$

$$\bar{x} = \frac{559}{108}$$

$$\bar{x} \approx 5.18$$

To find the median, list contracts cumulatively:

Number of Bidders	Number of Contracts	Cumulative Contracts
2	7	7
3	20	27
4	26	53
5	16	69
6	11	80
7	9	89
8	6	95
9	8	103
10	3	106
11	2	108

The median corresponds to the $\frac{108}{2} = 54$ th contract.

Cumulative up to 4 bidders is 53 (<54)

Up to 5 bidders is 69 (>54)

Thus, median number of bidders per contract is 5.

Mode is the number of bidders with the highest frequency (number of contracts).

Number of contracts per bidders:

Bidders	Contracts
4	26

The highest frequency is 26 at 4 bidders, so:

Mode = 4 bidders

Comparison of the three summaries:

We observe Mode < Median < Mean. This ordering (mean slightly larger than median, median larger than mode) indicates a mild right skew in the distribution.

b. Among the three numerical summaries:
- The mode was the simplest to determine, since it only required identifying the value (4 bidders) with the highest frequency.
- The mean required the most work, as it involved multiplying each bidder value by its frequency, adding these products, and dividing by the total number of contracts.
- The median fell in between: it needed cumulative frequencies to locate the middle observations but involved less computation than the mean.

48.

Solutions:

a. The sample mean will be higher than the sample median because the histogram is right-skewed, the mean is pulled upward by the few runners with very large time differences.

b. The sample median provides a value that indicates a difference in running times similar to a greater percentage of runners. Because the data is skewed, the mean is pulled higher by a few runners who slowed down a lot. The median isn't affected by those extreme cases, so it better represents the typical runner.

49.

Solutions:

a. The geometric mean for the given data will be $\sqrt[4]{1\times 5\times 19\times 3}$ which is equal to 4.11.

b. The sample mean $\bar{x}=\frac{28}{4}=7$ and sample median $\tilde{x}$ = average of 2nd and 3rd term after arranging = 4.

c. The geometric mean is more resistant to outliers than the sample mean. The large value (19) pulls the sample mean up to 7, but the geometric mean stays lower at about 4.18—closer to the typical values. So, the geometric mean gives a better sense of the center when there's an unusually large number.

SECTION 1.4

50.

Solution:

$$\sum_{i=1}^{n+1}x_i=\sum_{i=1}^{n}x_i+x_{n+1}=n\bar{x}_n+x_{n+1},\ \text{so}\ \bar{x}_{n+1}=\frac{1}{n+1}\sum_{i=1}^{n+1}x_i=\frac{n\bar{x}_n+x_{n+1}}{n+1}.$$

51.

Solutions:

a. $\bar{x} = 115.58$. The deviations from the mean are $116.4 - 115.58 = .82, 115.9 - 115.58 = .32$, $114.6 - 115.58 = -.98, 115.2 - 115.58 = -.38$, and $115.8 - 115.58 = .22$. Notice that the deviations from the mean sum to zero, as they should.

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

c. $\sum x_i^2 = 66795.61$, so $s^2 = s_{xx}/(n-1) = \left(\sum x_i^2 - \sum x_i)^2/\text{n}\right)/(n-1) =$ $(66795.61 - (577.9)^2/5)/4 = 1.928/4 = .482$.

d. The new sample values are: 16.4 15.9 14.6 15.2 15.8. While the new mean is 15.58, all the deviations are the same as in part (a), and the variance of the transformed data is identical to that of part (b).

52.

Solutions:

a. Range = Max – Min = 138.4 – 108.3 = 30.1. To find the standard deviation, we need to find the mean. The mean is $\bar{x} = \dfrac{118.6 + 127.4 + 138.4 + 130.0 + 113.7 + 122.0 + 108.3 + 131.5 + 133.2}{9} = 124.79$. The standard deviation is $s = \sqrt{\sum_{i=1}^{9} \dfrac{\left(x_i - \bar{x}\right)^2}{9-1}} = 9.84$. The fourth spread is the difference between the third and first quartiles. $Q1 = \dfrac{113.7 + 118.6}{2} = 116.15$ and $Q3 = \dfrac{131.5 + 133.2}{2} = 132.25$. So, the fourth spread is $Q3 - Q1 = 132.25 - 116.15 = 16.20$.

b. Range = 140 – 110 = 30. The mean is $\bar{x} = \dfrac{120 + 125 + 140 + 130 + 115 + 120 + 110 + 130 + 135}{9} = 125$. The The standard deviation is $s = \sqrt{\sum_{i=1}^{9} \dfrac{\left(x_i - \bar{x}\right)^2}{9-1}} = 9.68$. The fourth spread is the difference between the third and first quartiles. $Q1 = \dfrac{115 + 120}{2} = 117.5$ and $Q3 = \dfrac{130 + 135}{2} = 132.5$. So, the fourth spread is $Q3 - Q1 = 132.5 - 117.5 = 15$.

c. The range in part b is slightly less than part a and the standard deviation and fourth spread are also less in part (b) compared to part (a)

53.

Solutions:

a. The range of the data is $\text{Range} = \max - \min = 1285 - 350 = 935\,(1000\text{s }\$)$. The mean is $\bar{x} = \dfrac{\text{sum of the data}}{n} = 640.5\,(1000\text{s }\$)$. The standard deviation is, $s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})}{n-1}} = \sqrt{\dfrac{611066.5}{10-1}} == \sqrt{67896.277\ldots} \approx 260.569$ (1000s \$) and the fourth spread $f_s = \text{upper fourth} - \text{lower fourth} = 679 - 540 = 139$.

b. The range of the data is $\text{Range} = \max - \min = 985 - 350 = 635\,(1000\text{s }\$)$. The mean is $\bar{x} = \dfrac{\text{sum of the data}}{n} = 610.5\,(1000\text{s }\$)$. The standard deviation is, $s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})}{n-1}} = \sqrt{\dfrac{305366.5}{10-1}} = \sqrt{33929.6111\ldots} \approx 184.20\ (1000\text{s }\$)$ and the fourth spread f_s will remain the same as 139.

c. The data in part (a) had an outlier, 1285, which made both the range and the standard deviation larger. When this value was replaced with a smaller number, 985, both the range and the standard deviation decreased, so the data became less spread out. However, the fourth spread stayed the same because it only depends on the difference between the medians of the lower and upper fourths, which are not affected by a single outlier.

54.

Solutions:

a. Since all three distributions are somewhat skewed and two contain outliers (see **d**), medians are the more appropriate central measures. The medians are
Cooler: 1.760°C Control: 1.900°C Warmer: 2.305°C
The median difference between air and soil temperature increases as the conditions of the minichambers transition from cooler to warmer $(1.76 < 1.9 < 2.305)$.

b. With the aid of software, the standard deviations are
Cooler: 0.401°C Control: 0.531°C Warmer: 0.778°C
For the 15 observations under the "cooler" conditions, the typical deviation between an observed temperature difference and the mean temperature difference (1.760°C) is roughly 0.4°C. A similar interpretation applies to the other two standard deviations. We see that, according to the standard deviations, variability increases as the conditions of the minichambers transition from cooler to warmer $(0.401 < 0.531 < 0.778)$.

c. Apply the definitions of lower fourth, upper fourth, and fourth spread to the sorted data within each condition.
Cooler: lower fourth $= (1.43 + 1.57)/2 = 1.50$, upper fourth $= (1.88 + 1.90)/2 = 1.89$,

$f_s = 1.89 - 1.50 = 0.39°C$

Control: lower fourth $= (1.52 + 1.78)/2 = 1.65$, upper fourth $= (2.00 + 2.03)/2 = 2.015$,

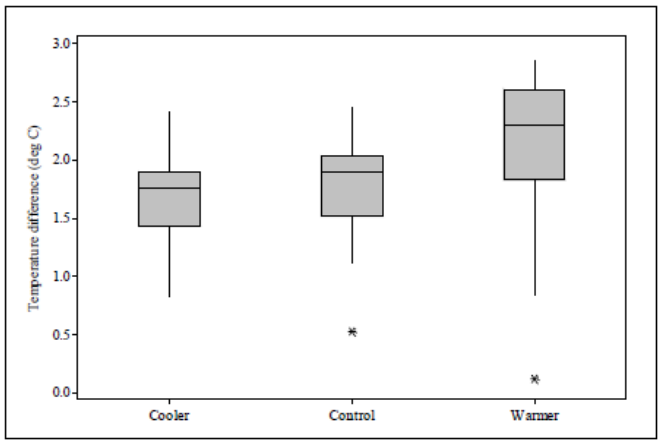
$f_s = 2.015 - 1.65 = 0.365°C$

Warmer: lower fourth $= 1.91$, upper fourth $= 2.60$,

$f_s = 2.60 - 1.91 = 0.69°C$

The fourth spreads do not communicate the same message as the standard deviations did. The fourth spreads indicate that variability is quite similar under the cooler and control settings, while variability is much larger under the warmer setting. The disparity between the results of **b** and **c** can be partly attributed to the skewness and outliers in the data, which unduly affect the standard deviations.

d. As noted earlier, the temperature difference distributions are negatively skewed under all three conditions. The control and warmer data sets each have a single outlier. The boxplots confirm that median temperature difference increases as we transition from cooler to warmer, that cooler and control variability are similar, and that variability under the warmer condition is quite a bit larger.



55.

Solutions:

a. From software, $\bar{x} = 14.7\%$ and $\bar{x} = 14.88\%$. The sample average alcohol content of these 10 wines was 14.88%. Half the wines have alcohol content below 14.7% and half are above 14.7% alcohol.

b. Working long-hand, $\sum (x_i - \bar{x})^2 = (14.8 - 14.88)^2 + \cdots + (15.0 - 14.88)^2 = 7.536$. The sample variance equals $s^2 = \sum (x_i - \bar{x})^2 = 7.536/(10 - 1) = 0.837$.

c. Subtracting 13 from each value will not affect the variance. The 10 new observations are 1.8, 1.5, 3.1, 1.2, 2.9, 0.7, 3.2, 1.6, 0.8, and 2.0. The sum and sum of squares of these 10 new numbers are $\sum y_i = 18.8$ and $\sum y_i^2 = 42.88$. Using the sample variance shortcut, we obtain $s^2 = \left[42.88 - (18.8)^2/10\right]/(10 - 1) = 7.536/9 =$ 0.837 again.

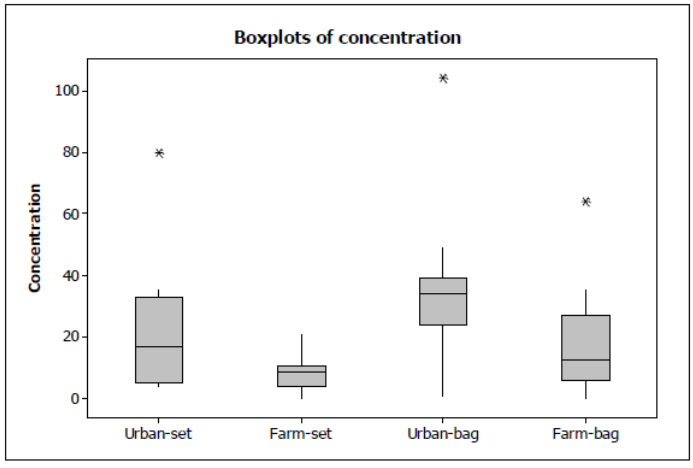
56.

Solutions:

a. Using the sums provided for urban homes, $S_{xx} = 10{,}079 - (237.0)^2/11 = 4972.73$, so $s = \sqrt{\frac{4972.73}{11-1}} = 22.3$ EU/mg. Similarly for farm homes, $S_{xx} = 518.36$ and $s = 6.09$ EU/mg. The endotoxin concentration in an urban home "typically" deviates from the average of 21.55 by about 22.3 EU/mg. The endotoxin concentration in a farm home "typically" deviates from the average of 8.56 by about 6.09 EU/mg. (These interpretations are very loose, especially since the distributions are not symmetric.) In any case, the variability in endotoxin concentration is much more in urban homes than in farm homes.

b. For urban data, the upper and lower fourths are 28.0 and 5.5, respectively, for a fourth spread of 22.5 EU/mg. For farm data, the upper and lower fourths are 10.1 and 4, respectively, for a fourth spread of 6.1 EU/mg. Again, we see that the variability in endotoxin concentration is much greater for urban homes than for farm homes.

c. Consider the box plots below. The endotoxin concentration in urban homes generally exceeds that in farm homes, whether measured by settled dust or bag dust. The endotoxin concentration in bag dust generally exceeds that of settled dust, both in urban homes and in farm homes. Settled dust in farm homes shows far less variability than any other scenario.



57.

Solutions:

a. $\sum x_i = 2.75 + \cdots + 3.01 = 56.80$, $\sum x_i^2 = 2.75^2 + \cdots + 3.01^2 = 197.8040$

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

58.

Solution:

From software or from the sums provided, $\bar{x} = 20179 / 27 = 747.37$ and $s = \sqrt{\dfrac{24657511 - (20179)^2 / 27}{26}} = 606.89$. The maximum award should be $\bar{x} + 2s = 747.37 + 2(606.89) = 1961.16$, or \$1,961,160. This is quite a bit less than the \$3.5 million that was awarded originally.

59.

Solutions:

a. From software, $s^2 = 1264.77 \text{ min}^2$ and $s = 35.56$ min. Working by hand, $\sum x = 2563$ and $\sum x^2 = 368501$, so

$$s^2 = \frac{368501 - (2563)^2/19}{19 - 1} = 1264.766 \text{ and } s = \sqrt{1264.766} = 35.564$$

b. If $y =$ time in hours, then $y = cx$ where $c = \dfrac{1}{60}$. So, $s_y^2 = c^2 s_x^2 = \left(\frac{1}{60}\right)^2$ $1264.766 = .351 \text{ hr}^2$ and $s_y = CS_x = \left(\dfrac{1}{60}\right) 35.564 = .593$ hr.

60.

Solution:

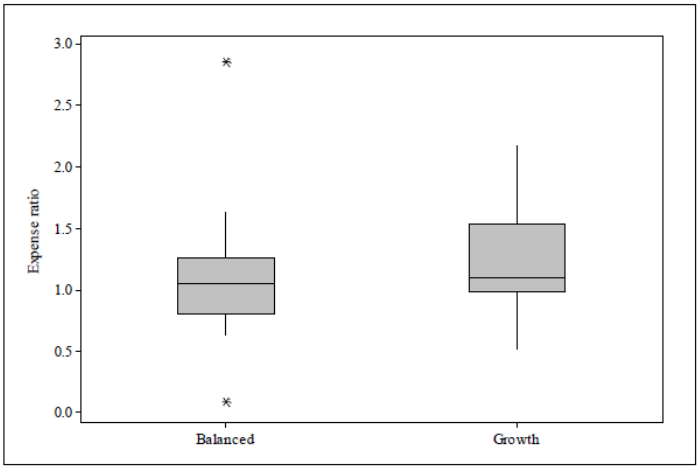
Let d denote the fifth deviation. Then $.3 + .9 + 1.0 + 1.3 + d = 0$ or $3.5 + d = 0$, so $d = -3.5$. One sample for which these are the deviations is $x_1 = 3.8$, $x_2 = 4.4$, $x_3 = 4.5$, $x_4 = 4.8$, $x_5 = 0$. (These were obtained by adding 3.5 to each deviation; adding any other number will produce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

61.

Solutions:

a. Balance Fund, forth spread = 0.45
Growth Fund, forth spread = 0.535

b. Typical ratios are quite similar for the two types. There is somewhat more variability in the Bal sample, due primarily to the two outliers (one mild, one extreme). For Bal, there is substantial symmetry in the middle 50% but positive skewness overall. For Gr, there is substantial positive skew in the middle 50% and mild positive skewness overall.



62.

Solutions:

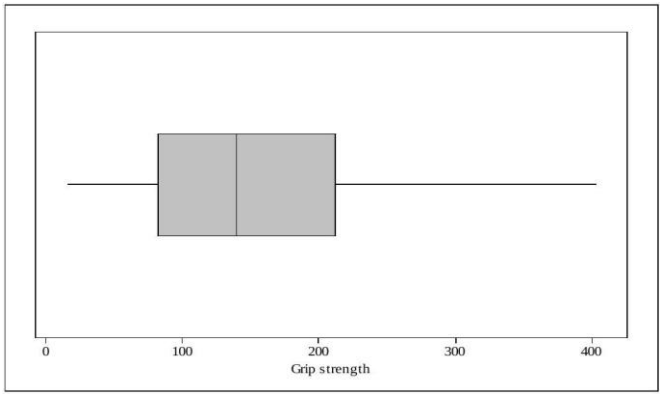
a. The stem-and-leaf display provided by Minitab is shown below. Grip strengths for this sample of 42 individuals are positively skewed, and there is one high outlier at 403 N.

6	0	111234	
14	0	55668999	
(10)	1	0011223444	Stem = 100s
18	1	567889	Leaf = 10s
12	2	01223334	
4	2	59	
2	3	2	
1	3		
1	4	0	

b. Each half has 21 observations. The lower fourth is the 11th observation, 87 N. The upper fourth is the 32nd observation (11th from the top), 210 N. The fourth spread is the difference: $f_s = 210 - 87 = 123$ N.

c. min = 16; lower fourth = 87; median = 140; upper fourth = 210; max = 403

The boxplot tells a similar story: grip strengths are slightly positively skewed, with a median of 140 N and a fourth spread of 123 N.



d. inner fences: $87 - 1.5(123) = -97.5,\ 210 = 394.5$

outer fences: $87 - 3(123) = -282,\ 210 + 3(123) = 579$

Grip strength can't be negative, so low outliers are impossible here. A mild high outlier is above 394.5 N and an extreme high outlier is above 579 N. The value 403 N is a mild outlier by this criterion. (Note: some software uses slightly different rules to define outliers — using quartiles and interquartile range — which result in 403 N not being classified as an outlier.)

e. The fourth spread is unaffected unless that observation drops below the current upper fourth, 210. That's a decrease of $403 - 210 = 193$ N.

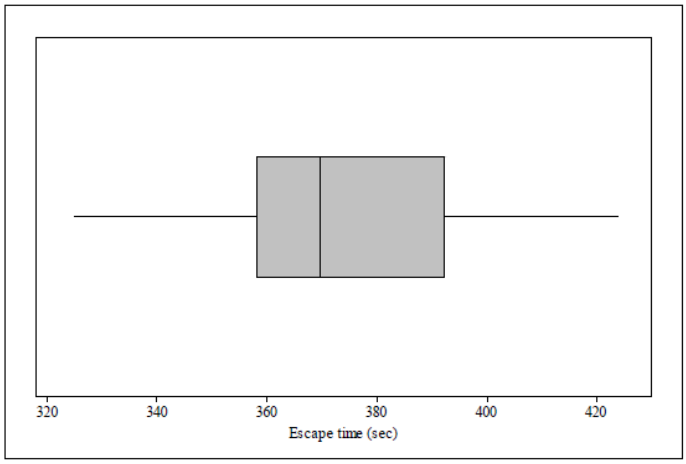
63.

Solutions:

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 fourth, is 359 (the 7^{th} observation in the sorted list).
Upper half of the data set: 370 373 373 374 375 389 392 393 394 397 402 403 424, whose median, and therefore the upper fourth is 392. So, $f_s = 392 - 359 = 33$.

b. inner fences: $359 - 1.5(33) = 309.5, 392 + 1.5(33) = 441.5$
To be a mild outlier, an observation must be below 309.5 or above 441.5. There are none in this data set. Clearly, then, there are also no extreme outliers.

c. A boxplot of this data appears below. The distribution of escape times is roughly symmetric with no outliers. Notice the box plot "hides" the fact that the distribution contains two gaps, which can be seen in the stem-and-leaf plot.

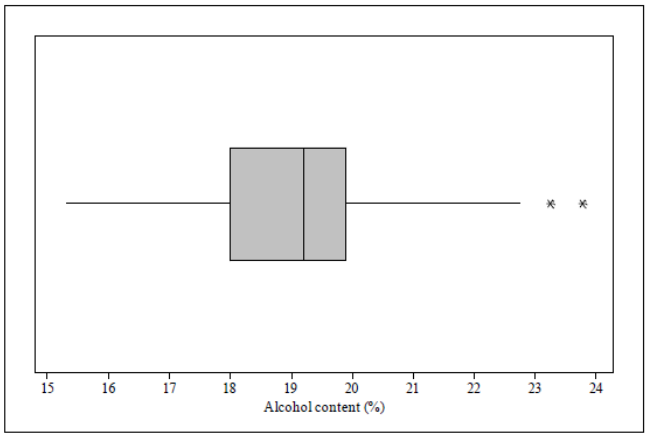


d. Not until the value $x = 424$ is lowered below the upper fourth value of 392 would there be any change in the value of the upper fourth (and, thus, of the fourth spread). That is, the value $x = 424$ could not be decreased by more than $424 - 392 = 32$ seconds.

64.

Solution:

The alcohol content distribution of this sample of 35 port wines is roughly symmetric except for two high outliers. The median alcohol content is 19.2% and the fourth spread is 1.42%. [upper fourth = (19.90 + 19.62)/2 = 19.76; lower fourth = (18.00 + 18.68)/2 = 18.34] The two outliers were 23.25% and 23.78%, indicating two port wines with unusually high alcohol content.



65.

Solutions:

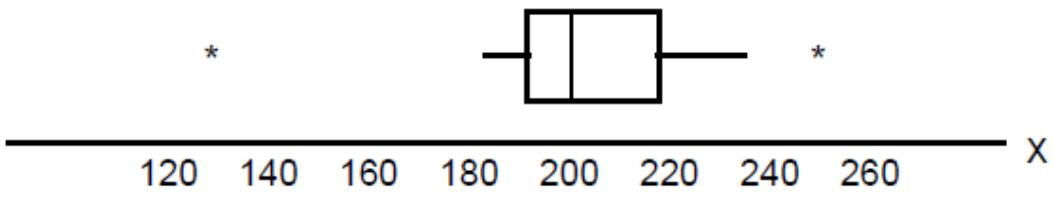
a. $f_s = 216.8 - 196.0 = 20.8$

inner fences: $196 - 1.5(20.8) = 164.6$, $216.8 + 1.5(20.8) = 248$

outer fences: $196 - 3(20.8) = 133.6$, $216.8 + 3(20.8) = 279.2$

Of the observations listed, 125.8 is an extreme low outlier and 250.2 is a mild high 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.



66.

Solutions:

a. 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.

b. 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.

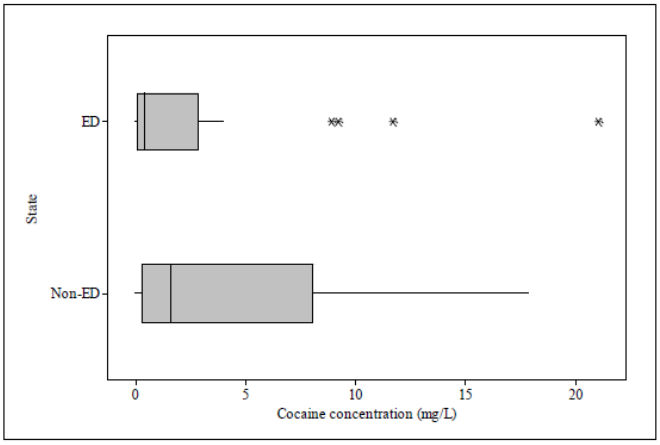
67.

Solutions:

a. If you aren't using software, don't forget to sort the data first!
ED: median = .4, lower fourth = (.1 + .1)/2 = .1, upper fourth = (2.7 + 2.8)/2 = 2.75,
fourth spread = 2.75 – .1 = 2.65,
Non-ED: median = (1.5 + 1.7)/2 = 1.6, lower fourth = .3, upper fourth = 7.9,
fourth spread = 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 cannot be any extreme outliers, either.

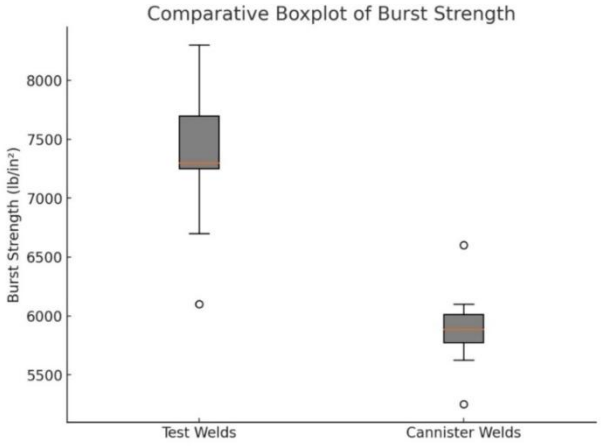
c. A comparative boxplot is shown below. The outliers in the ED data are clearly visible. There is noticeable positive skewness in both samples. The Non-ED sample has more variability than the Ed sample and the typical values of the ED sample tend to be smaller than those for the Non-ED sample.



68.

Solutions:

a. The Test Welds have a higher median burst strength than Cannister Welds. The spread of the Test Welds data is larger, showing more variability.



b. Test Welds:
Mean = 7354.55 lb/in²
Standard Deviation = 613.78 lb/in²
Cannister Welds:
Mean = 5887.5 lb/in²
Standard Deviation = 317.93 lb/in²

c. The boxplot is more useful for understanding distribution differences because it visually displays the spread, skewness, and outliers.

69.

Solutions:

a. Description of Each Distribution
   1. 6 a.m.
      - The median is relatively low, closer to 7-8.
      - The interquartile range (IQR) is relatively small.
      - There are multiple outliers, including a very extreme one above 60.
      - The whiskers are fairly long, indicating some spread in the data.
   2. 8 a.m.
      - The median is slightly higher than at 6 a.m.
      - The IQR is similar to that of 6 a.m.
      - The whiskers are slightly shorter compared to 6 a.m.
   3. 12 noon
      - The median has increased to 10 a.m.
      - The IQR is larger than the previous two time points.
      - The whiskers are longer, indicating more spread in the data.
      - There are fewer outliers.
   4. 2 p.m.
      - The median is at its highest among all time points.
      - The IQR is the largest among the five distributions.

- The whiskers are quite long, indicating a wide range of values.
- There are very few or no visible outliers.

5. 10 p.m.
   - The median has decreased compared to 2 p.m.
   - The IQR is slightly smaller than at 2 p.m. but still relatively large.
   - There are no major outliers.

b. Comparison and Observations

- The median gas vapor coefficient increases from morning (6 a.m.) to afternoon (2 p.m.), then decreases by 10 p.m. This shows that gasoline vapor levels are highest in the early afternoon.
- Variability (IQR) increases from 6 a.m. to 2 p.m., indicating that gasoline vapor coefficients fluctuate more during the day.
- The 6 a.m. data has the most extreme outliers, indicating that some vehicles emit much higher levels in the early morning.
- The whiskers are longest at 2 p.m., showing that gasoline vapor coefficients are most spread out at this time.

70.

Solution:

Given: $n = 4, x_1 = 76683, x_4 = 77048, \bar{x} = 76831, s = 180.$

Goal: find the two middle observations x_2 and x_3.

Sum constraint (from the mean)

$$x_1 + x_2 + x_3 + x_4 = 4\bar{x} \Rightarrow x_2 + x_3 = 4\bar{x} - (x_1 + x_4) = 4(76831) - (76683 + 77048) = 153593 \ldots (1).$$

Sum-of-squares constraint (from the sample variance):

$S_{xx} = (n-1)s^2 = 97200 \Rightarrow \sum_{i=1}^{4} x_i^2 = S_{xx} + n\bar{x}^2 = 23612107444$. So,

$$x_2^2 + x_3^2 = 23612107444 - x_1^2 - x_4^2 = 11795430651 \ldots (2)$$

Let $S = x_2 + x_3$ and $P = x_2 x_3$. From (1) and (2),

$$x_2^2 + x_3^2 = S^2 - 2P \Rightarrow P = \frac{S^2 - (x_2^2 + x_3^2)}{2} = \frac{153593^2 - 11795430651}{2} = 5897689499.$$

Solve the quadratic $t^2 - St + P = 0$: $t = \frac{S \pm \sqrt{\Delta}}{2}$,

$\Delta = S^2 - 4P \approx 51653 \Rightarrow \sqrt{\Delta} \approx 227.2729636$.

$$t_1 \approx \frac{153593 + 227.2729636}{2} \approx 76910$$

$$t_2 \approx \frac{153593 - 227.2729636}{2} \approx 76683$$

Assign in increasing order $x_1 \le x_2 \le x_3 \le x_4$. The two middle observations are $x_2 \approx 76683,\ x_3 \approx 76910$

71.

Solution:

Since the exact maximum value(outlier) is not provided to us, so we cannot calculate range, variance or the standard deviation. Only fourth spread can be calculated from this data as it requires the difference between the median of lower and upper fourth making it having no need for an exact value of an outlier. The value of fourth spread $f_s = 92 - 35 = 57$.

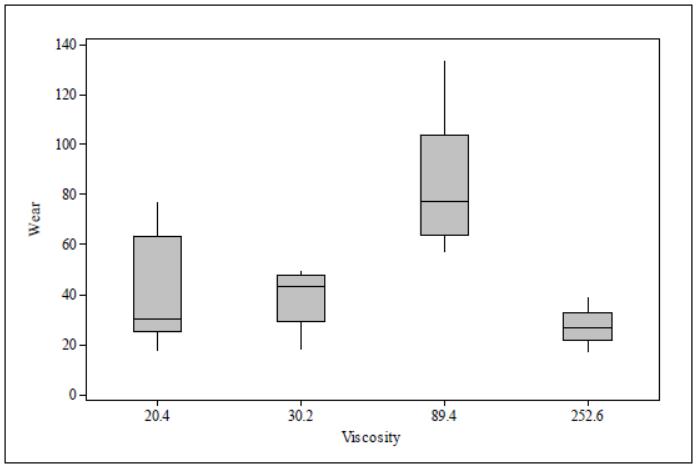
72.

Solutions:

a. The sample coefficient of variation is similar for the three highest oil viscosity levels (29.66, 32.30, 27.86) but is much higher for the lowest viscosity (56.01). At low viscosity, it appears that there is much more variation in volume wear relative to the average or "typical" amount of wear.

	Wear		
Viscosity	$\bar{x}$	s	cv
20.4	40.17	22.50	56.01
30.2	38.83	11.52	29.66
89.4	84.10	27.20	32.30
252.6	27.10	7.55	27.86

b. Volume wear changes a lot depending on viscosity. At very high viscosity, wear is typically the least and the least variable. Volume wear is actually by far the highest at a medium viscosity level and also has the greatest variability at this viscosity level. Lower viscosity levels lead to less wear than medium, but very low viscosity shows much more relative variation.



73.

Solution:

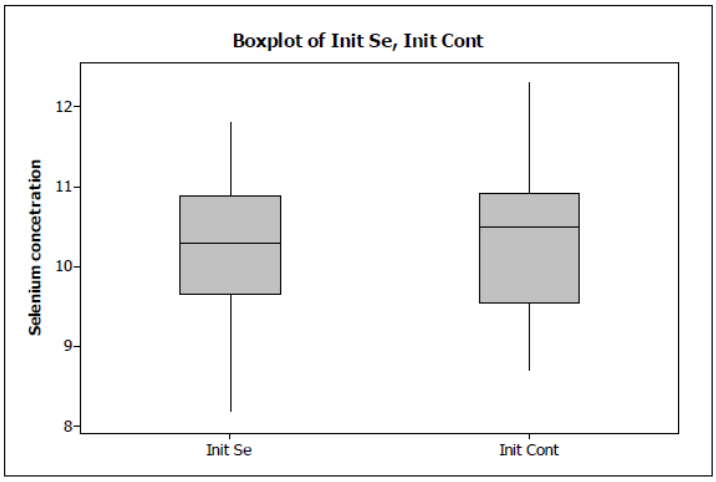
Assuming that the histogram is unimodal, then there is evidence of positive skewness in the data since the median lies to the left of the mean (for a symmetric distribution, the mean and median would coincide).

For more evidence of skewness, compare the distances of the 5^{th} and 95^{th} percentiles from the median: median -5^{th} %ile $= 500 - 400 = 100$, while 95^{th} %ile – 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 10^{th} and 90^{th} percentiles to the median, or comparing the maximum and minimum to the median.

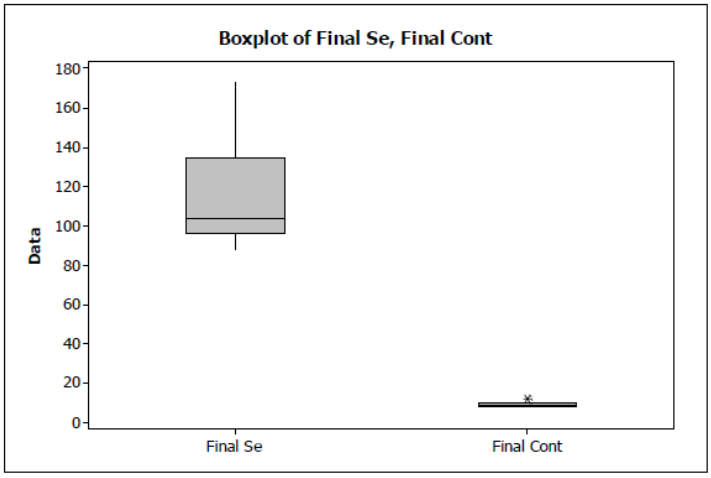
74.

Solutions:

a. The initial Se concentrations in the treatment and control groups are very similar. The box plots below shown only minor differences. The median initial Se concentrations for the treatment and control groups are 10.3 mg/L and 10.5 mg/L, respectively, each with fourth spread of about 1.25 mg/L. So, the two groups of cows are comparable at the beginning of the study.



b. The final Se concentrations of the two groups are extremely different, as seen in the box plots below. Whereas the median final Se concentration for the control group is 9.3 mg/L (actually slightly lower than the initial concentration), the median Se concentration in the treatment group is now 103.9 mg/L, nearly a 10-fold increase.



75.

Solutions:

a. Aortic root diameters for males have mean 3.6 cm, median 3.70 cm, standard deviation 0.2914 cm, and fourth spread 0.50. The corresponding values for females are $\bar{x} = 3.28$ cm, $\tilde{x} = 3.15$ cm, $s = 0.478$ cm, and $f_s = 0.50$ cm. Aortic root diameters are typically (though not universally) somewhat smaller for females than for males, and females show more variability.

b. For females $(n = 10)$, the 10% trimmed mean is the average of the middle 8 observations: $\bar{x}_{tr(10)} = 3.69$ cm For males $(n = 13)$, the 1/13 trimmed mean is $40.2/11 = 3.6545$, and the 2/13 trimmed mean is $32.8/9 = 3.6444$. Interpolating, the 10% trimmed mean is $\bar{x}_{tr(10)} = 0.7(3.6545) + 0.3(3.6444) = 3.65$ cm. (10% is three-tenths of the way from 1/13 to 2/13).

76.

Solutions:

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

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

b. Since $c = \bar{x}$ minimizes $\sum(x_i - c)^2$, $\sum(x_i - \bar{x})^2 < \sum(x_i - \mu)^2$.

77.

Solutions:

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

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

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

b. $x = °\text{C}$, $y = °\text{F}$

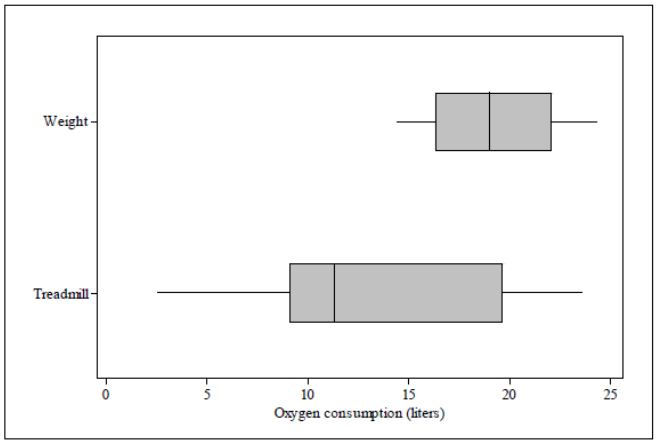
$$\bar{y} = \frac{9}{5}(87.3) + 32 = 189.14°\text{F}$$

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

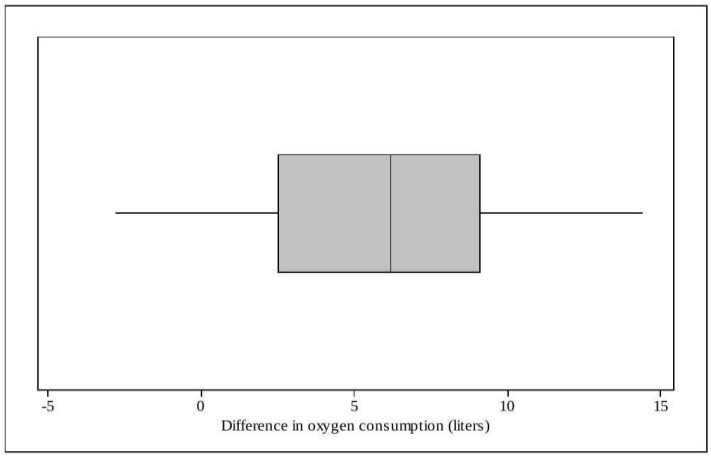
78.

Solutions:

a. There is a significant difference in the variability of the two samples. On average, the weight training led to much higher oxygen consumption, than the treadmill exercise, with the median values around 20 and 11 liters, respectively.



b. The differences in oxygen consumption (weight minus treadmill) for the 15 subjects 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 than the treadmill did. The median difference is about 6 liters.

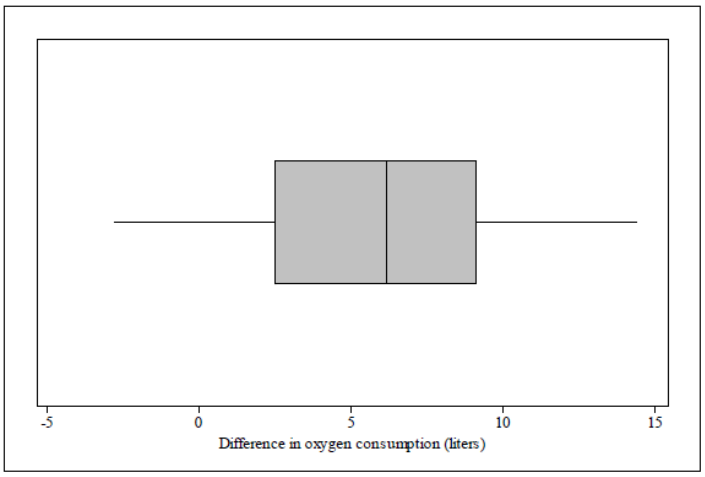


79.

Solutions:

a. The mean is 135.39. The standard deviation is 4.59. The five number summary is (in order) 122.20, 132.95, 135.40,138.25, 147.70.

b. Refer to the comments for (a). In addition, using $1.5(Q3-Q1)$ as a yardstick, the two largest and three smallest observations are mild outliers.



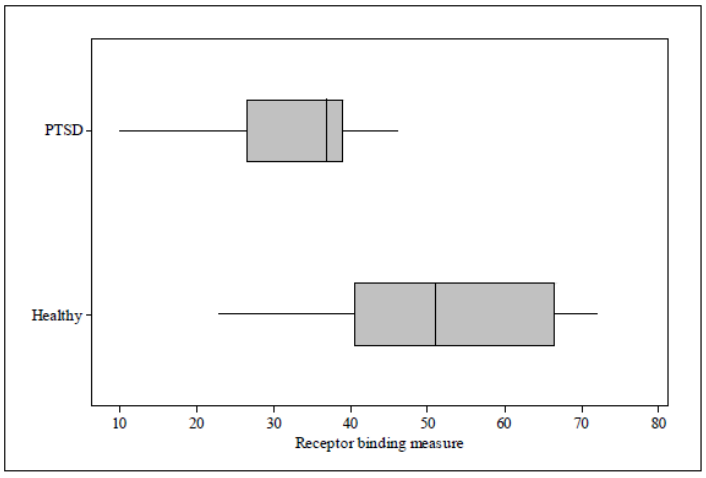
80.

Solution:

A table of summary statistics, a stem and leaf display, and a comparative boxplot are shown below. On average, the healthy individuals have higher receptor binding measure than the individuals with PTSD. There is also more variation in the healthy individuals' values. The distribution of values for the healthy individual is reasonably symmetric, while the distribution for the PTSD individuals is negatively skewed. The box plot indicates that there are no outliers, and supports the observations about 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

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

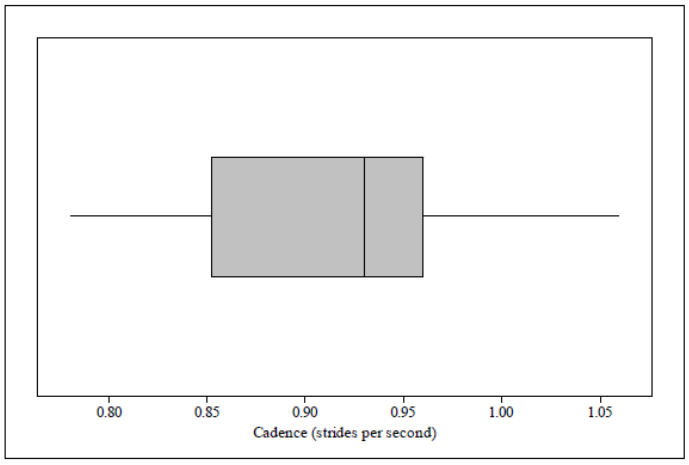


81.

Solution:

From software, $\bar{x} = .9255$, $s = .0809$; $\tilde{x} = .93$, $f_s = .1$. The cadence observations are slightly (right) skewed (mean = .9255 strides/sec, median = .93 strides/sec) and show a small amount of variability (standard deviation = .0809, fourth spread = .1). There are no apparent outliers in the data.

7	8	stem = tenths
8	11556	leaf = hundredths
9	2233335566	
0	0566	



82.

Solution:

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

83.

Solutions:

a. R code for Stem and leaf plot:

```
depth <- c(0.41, 0.41, 0.41, 0.41, 0.43, 0.43, 0.43, 0.48, 0.48,
+          0.58, 0.79, 0.79, 0.81, 0.81, 0.81, 0.91, 0.94, 0.94,
+          1.02, 1.04, 1.04, 1.17, 1.17, 1.17, 1.17, 1.17, 1.17,
+          1.17, 1.19, 1.19, 1.27, 1.40, 1.40, 1.59, 1.59, 1.60,
+          1.68, 1.91, 1.96, 1.96, 1.96, 2.10, 2.21, 2.31, 2.46,
+          2.49, 2.57, 2.74, 3.10, 3.18, 3.30, 3.58, 3.58, 4.15,
+          4.75, 5.33, 7.65, 7.70, 8.13, 10.41, 13.44)
>
> stem(depth, scale = 2)
```

Plot:

```
 0 | 444444455688888999
 1 | 00022222222234466679
 2 | 0001235567
 3 | 12366
 4 | 28
 5 | 3
 6 |
 7 | 77
 8 | 1
 9 |
10 | 4
11 |
12 |
13 | 4
```

b. Representative depths are quite similar for the three types of soils — between 1.5 and 2. Data from the C and CL soils shows much more variability than for the other two types. The boxplots for the first three types show substantial positive skewness both in the middle 50% and overall. The boxplot for the SYCL soil shows negative skewness in the middle 50% and mild positive skewness overall. Finally, there are multiple outliers for the first three types of soils, including extreme outliers.

84.

Solutions:

a. Since the constant $\bar{x}$ 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 part (a), of the y's). Then $z_i = cy_i \Rightarrow s_z^2 = c^2 s_y^2 = (1/s)^2 s^2 = 1$ and $s_z = 1$. That is, the "standardized" quantities $z_1, \ldots, z_n$ have a sample variance and standard deviation of 1.

85.

Solution:

We artificially add and subtract $n\bar{x}_n^2$ to create the term needed for the sample variance:

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

$$= \sum_{i=1}^{n} x_i^2 + x_{n+1}^2 - (n+1)\bar{x}_{n+1}^2 = \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2\right] + n\bar{x}_n^2 + x_{n+1}^2 - (n+1)\bar{x}_{n+1}^2 \text{ n}$$

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

$$= (n-1)s_n^2 + x_{n+1}^2 + n\bar{x}_n^2 - (n+1)\left[\frac{n}{n+1}\bar{x}_n + \frac{x_{n+1}}{n+1}\right]^2$$

$$= (n-1)s_n^2 + x_{n+1}^2 + n\bar{x}_n^2 - \left[\frac{(n\bar{x}_n + x_{n+1})^2}{n+1}\right]$$

$$= (n-1)s_n^2 + x_{n+1}^2 + n\bar{x}_n^2 - \left[\frac{n^2\bar{x}_n^2 + 2n\bar{x}_n x_{n+1} + x_{n+1}^2}{n+1}\right]$$

$$= (n-1)s_n^2 + x_{n+1}^2 - \frac{x_{n+1}^2}{n+1} + n\bar{x}_n^2 - \frac{n^2\bar{x}_n^2}{n+1} - \frac{2n\bar{x}_n x_{n+1}}{n+1}$$

$$= (n-1)s_n^2 + \frac{n}{n+1}x_{n+1}^2 + \frac{n(n+1)-n^2}{n+1}\bar{x}_n^2 - \frac{2n\bar{x}_n x_{n+1}}{n+1}$$

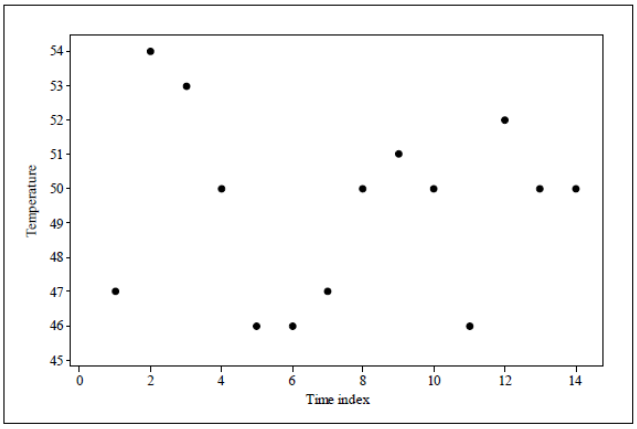
$$= (n-1)s_n^2 + \frac{n}{n+1}\left(x_{n+1}^2 - 2\bar{x}_n x_{n+1} + \bar{x}_n^2\right)$$

$$= (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2$$

86.

Solutions:

a. There is some evidence of a cyclical pattern.



b. A complete listing of the smoothed values appears below. To illustrate, with $\alpha = .1$ we have $\bar{x}_2 = .1x_2 + .9\bar{x}_1 = (.1)(54) + (.9)(47) = 47.7$, $\bar{x}_3 = .1x_3 + .9\bar{x}_2 = (.1)(53) + (.9)(47.7) = 48.23 \approx 48.2$, etc. It's clear from the values below that α = .1 gives a smoother series.

t	$\bar{x}_t$ for $\alpha = .1$	$\bar{x}_t$ for $\alpha = .5$
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

c. As seen below, $\bar{x}_t$ 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).

$$\begin{aligned}\bar{x}_t &= \alpha x_t + (1-\alpha)\bar{x}_{t-1} = \alpha x_t + (1-\alpha)\left[\alpha x_{t-1} + (1-\alpha)\bar{x}_{t-2}\right] \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2\left[\alpha x_{t-2} + (1-\alpha)\bar{x}_{t-3}\right] = \cdots \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots + \alpha(1-\alpha)^{t-2} x_2 + (1-\alpha)^{t-1} x_1\end{aligned}$$

d. For large t, the smoothed series is not very sensitive to the initial value x_1, since the coefficient $(1-\alpha)^{t-1}$ will be very small.

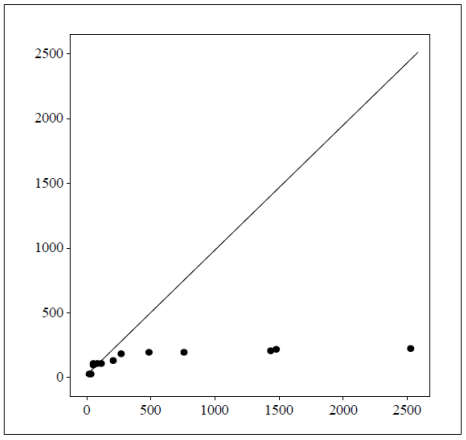
87.

Solutions:

a. When the data is perfectly symmetric, the plot of paired distances from the median forms a straight diagonal line going upward, where each point lies exactly on the line y = x. This means the distance from the median to the smallest value matches the distance from the median to the largest value, and so on for each pair. If the data has a long upper tail, meaning the values stretch farther above the median than below, then the points in the plot will fall above the diagonal line. That shows the upper half of the data is more spread out than the lower half. If the tail is longer on the lower side, the points fall below the line. The farther the points are from the diagonal, the more asymmetric the data is.

b. The median of these $n = 26$ observations is 221.6 (the midpoint of the 13th and 14th ordered values). The first point in the plot is $(2745.6 - 221.6,\ 221.6 - 4.1) =$ (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

exceeds the second number, so each of those points falls well below the 45° line. A substantial positive skew (stretched upper tail) is indicated.



88.

Solution:

As suggested in the hint, split the sum into two "halves" corresponding to the lowest $n/2$ observations and the highest $n/2$ observations (we'll use L and U to de note these).

$$\begin{aligned}\sum\left|x_i-\tilde{x}\right| &= \sum_L\left|x_i-\tilde{x}\right|+\sum_U\left|x_i-\tilde{x}\right| \\ &= \sum_L\left(\tilde{x}_i-x\right)+\sum_U\left(x_i-\tilde{x}\right) \\ &= \left|\sum_L\tilde{x}-\sum_L x_i+\sum_U x_i-\sum_U\tilde{x}\right|\end{aligned}$$

Each of these four sums covers exactly $n/2$ terms. The first and fourth sums are, therefore, both equal to $\left(n/2\right)\tilde{x}$; these cancel. The inner two sums may be re-written in terms of averages:

$$\begin{aligned}\sum\left|x_i-\tilde{x}\right| &= \sum_L\tilde{x}-\sum_L x_i+\sum_U x_i=-\sum_L x_i+\sum_U x_i \\ &= -\left(n/2\right)\bar{x}_L+\left(n/2\right)\bar{x}_U \quad \Rightarrow \\ \sum\left|x_i-\tilde{x}\right|/n &= \left(\bar{x}_U-\bar{x}_L\right)/2.\end{aligned}$$

There is a term missing from the top set of sums: $-\sum_U x_i$

When n is odd, the middle (ordered) value is exactly $\tilde{x}$. Using L and U to denote the lowest $\left(n-1\right)/2$ observations and largest $\left(n-1\right)/2$ observations, respectively, we may write $\sum\left|x_i-\tilde{x}\right|=\sum_L\left|x_i-\tilde{x}\right|+0+\sum_U\left|x_i-\tilde{x}\right|$, where the 0 comes from the middle (ordered) value, viz. $\left|\tilde{x}-\tilde{x}\right|=0$. The rest of the derivation proceeds exactly as before, except that the surviving sums each have $\left(n-1\right)/2$ terms in them, not $n/2$. As a result, for n odd we have $\sum\left|x_i-\tilde{x}\right|/\left(n-1\right)=\left(\bar{x}_U-\bar{x}_L\right)/2$.