Section 1.2

11. Every score in the following batch of exam scores is in the 60s, 70s, 80s, or 90s. A stem-and-leaf display with only the four stems 6, 7, 8, and 9 would not give a very detailed description of the distribution of scores. In such situations, it is desirable to use repeated stems. Here we could repeat the stem 6 twice, using 6L for scores in the low 60s (leaves 0, 1, 2, 3, and 4) and 6H for scores in the high 60s (leaves 5, 6, 7, 8, and 9). Similarly, the other stems can be repeated twice to obtain a display consisting of eight rows. Construct such a display for the given scores. What feature of the data is highlighted by this display?

74 89 80 93 64 67 72 70 66 85 89 81 81
71 74 82 85 63 72 81 81 95 84 81 80 70
69 66 60 83 85 98 84 68 90 82 69 72 87
88

Stem: low and high of 60s, 70s, 80s, 90s

leaf: the one digits. Low stem with the digit (0~4) high stem with (5~9)

11.

Stem	leaf
6L	4 3 0
6H	7 6 9 6 8 9
7L	4 2 0 1 4 2 0 2
7H	
8L	0 1 1 2 1 1 4 1 0 3 4 2
8H	9 5 9 5 5 7 8
9L	3 0
9H	5 8

from the stem-and-leaf display, we find that the display isn't highly concentrated, there is a gap at 7H (high 70's)

14. The accompanying data set consists of observations on shower-flow rate (L/min) for a sample of $n = 129$ houses in Perth, Australia ("An Application of Bayes Methodology to the Analysis of Diary Records in a Water Use Study," *J. Amer. Stat. Assoc.*, 1987: 705–711):

4.6 12.3 7.1 7.0 4.0 9.2 6.7 6.9 11.5 5.1
11.2 10.5 14.3 8.0 8.8 6.4 5.1 5.6 9.6 7.5
7.5 6.2 5.8 2.3 3.4 10.4 9.8 6.6 3.7 6.4
8.3 6.5 7.6 9.3 9.2 7.3 5.0 6.3 13.8 6.2
5.4 4.8 7.5 6.0 6.9 10.8 7.5 6.6 5.0 3.3
7.6 3.9 11.9 2.2 15.0 7.2 6.1 15.3 18.9 7.2
5.4 5.5 4.3 9.0 12.7 11.3 7.4 5.0 3.5 8.2
8.4 7.3 10.3 11.9 6.0 5.6 9.5 9.3 10.4 9.7
5.1 6.7 10.2 6.2 8.4 7.0 4.8 5.6 10.5 14.6
10.8 15.5 7.5 6.4 3.4 5.5 6.6 5.9 15.0 9.6
7.8 7.0 6.9 4.1 3.6 11.9 3.7 5.7 6.8 11.3
9.3 9.6 10.4 9.3 6.9 9.8 9.1 10.6 4.5 6.2
8.3 3.2 4.9 5.0 6.0 8.2 6.3 3.8 6.0

a. Construct a stem-and-leaf display of the data.
b. What is a typical, or representative, flow rate?
c. Does the display appear to be highly concentrated or spread out?
d. Does the distribution of values appear to be reasonably symmetric? If not, how would you describe the departure from symmetry?
e. Would you describe any observation as being far from the rest of the data (an outlier)?

14. Stem: tens and ones digits; leaf: one decimal place down.

(a)

stem	leaf
02	3 2
03	4 7 3 9 5 4 6 7 2 8
04	6 0 8 3 8 1 5 9
05	1 1 6 8 0 4 0 4 5 0 6 1 6 5 9 7 0
06	7 9 4 2 6 4 5 3 2 0 9 6 1 0 7 2 4 6 9 8 9 2 0 3 0
07	1 0 5 5 6 3 5 5 6 2 2 4 3 0 5 8 0
08	0 8 3 2 4 4 3 2
09	2 6 8 3 2 0 5 3 7 6 3 6 3 8 1
10	5 4 8 3 4 2 5 8 4 6
11	5 2 9 3 9 9 3
12	3 7
13	8
14	3 6
15	0 3 5 0
16	
17	
18	9

(b) from the stem-and-leaf display, the typical flow rate is the flow rate with the number of integer bits in 06. and the number from 6.9 to 7.0 is most concentrated (typical)

(c) the display isn't highly concentrated, it has gap since there is no data in 16 and 17 stem

(d) it isn't reasonably symmetric, it is positively skewed.

(e) there is an outlier: 18.9

Section 1.2

20. The ... Most Representative Subdivisions," ... 43–55) gave data on ... that could be used in deciding whether to provide electrical power using overhead lines or underground lines. Here are the values of the variable x = total length of streets within a subdivision:

1280	5320	4390	2100	1240	3060	4770
1050	360	3330	3380	340	1000	960
1320	530	3350	540	3870	1250	2400
960	1120	2130	450	2850	2320	2400
3150	5700	5220	500	1850	2460	5850
2700	2730	1670	100	5770	3150	1890
510	240	396	1410	2100		

a. Construct a stem-and-leaf display using the thousands digit as the stem and the hundreds digit as the leaf, and comment on the various features of the display.

b. Construct a histogram using class boundaries 0, 1000, 2000, 3000, 4000, 5000, and 6000. What proportion of subdivisions have total length less than 2000? Between 2000 and 4000? How would you describe the shape of the histogram?

20. Stem: thousands digits

leaf: hundreds digit (deleting the tens and ones digit

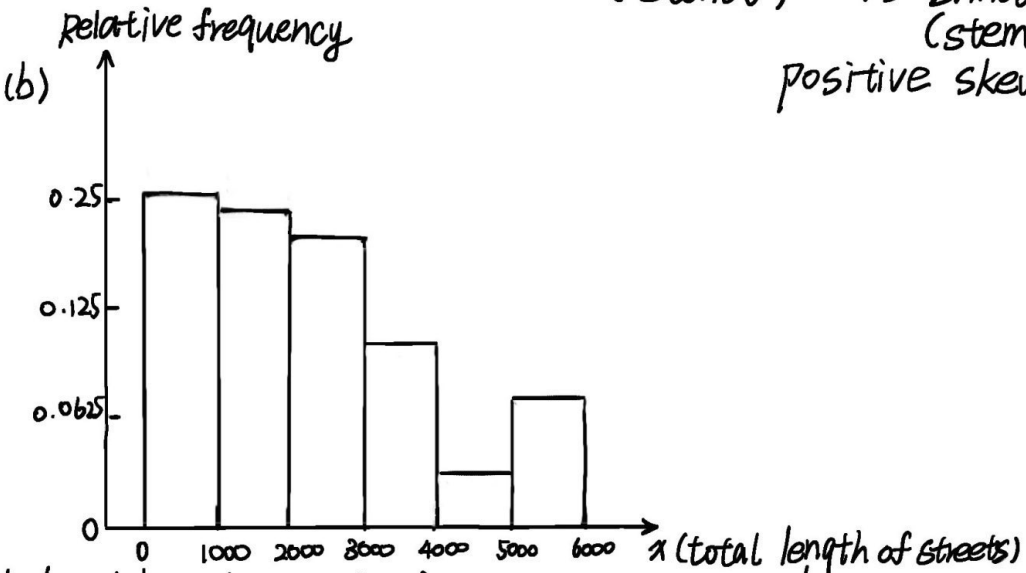
(a)

stem	leaf
0	3 3 9 5 5 9 4 5 1 5 2 3
1	2 2 0 0 3 2 1 8 6 8 4
2	1 4 1 2 [illegible] 4 7 7 1
3	0 3 3 3 8 1 1
4	3 7
5	3 7 2 8 7

since it's fairly evenly distributed, the typical data value is in low 2000, the display (stem 0) is bimodal (stem 0 and 5)
positive skew (?)

interval	number	frequency
0 – 1000	13	$\frac{13}{47}$
1000 – 2000	11	$\frac{11}{47}$
2000 – 3000	10	$\frac{10}{47}$
3000 – 4000	7	$\frac{7}{47}$
4000 – 5000	2	$\frac{2}{47}$
5000 – 6000	5	$\frac{5}{47}$

(b)



(b) the proportion of subdivisions have total length less than 2000 is $\frac{23}{47} \approx 0.489$

Between 2000 and 4000 : $\frac{17}{47} \approx 0.362$ the shape of the histogram is positively skewed

(Bimodal)

Section 1.3

...ially endotoxin, may ...llergic diseases. The ...for Endotoxin—An ... *(Indoor Air,* 2006: ...iated with determining endotoxin concentration. The following data on concentration (EU/mg) in settled dust for one sample of urban homes and another of farm homes was kindly supplied by the authors of the cited article.

	4	5	5	6	11	17	18	23	33	35	80		
U:	6.0	5.0	11.0	33.0	4.0	5.0	80.0	18.0	35.0	17.0	23.0		
F:	4.0	14.0	11.0	9.0	9.0	8.0	4.0	20.0	5.0	8.9	21.0	9.2	3.0	2.0	0.3

a. Determine the sample mean for each sample. How do they compare?

b. Determine the sample median for each sample. How do they compare? Why is the median for the urban sample so different from the mean for that sample?

c. Calculate the trimmed mean for each sample by deleting the smallest and largest observation. What are the corresponding trimming percentages? How do the values of these trimmed means compare to the corresponding means and medians?

24.

(a) the sample mean:

U: $\bar{X}_u = \frac{6.0+5.0+11.0+33.0+4.0+5.0+80.0+18.0+35.0+17.0+23.0}{11}$

$= \frac{237}{11} \approx 21.55$ (EU/mg)

F: $\bar{X}_F = \frac{4.0+14.0+11.0+9.0+9.0+8.0+4.0+20.0+5.0+8.9+21.0+9.2+3.0+2.0+0.3}{15}$

$= \frac{128.4}{15} \approx 8.56$ (EU/mg) Compare with two sample mean, we find the mean of concentration in settled dust in urban homes is higher than in farm homes.

(b) the sample median: U: $\tilde{X}_u = 17$ (EU/mg)

F: $\tilde{X}_F = 8.9$ (EU/mg)

the sample median in urban home higher than (of endotoxin concentration) farm home (nearly double)

since the large value is less in urban homes, and it has the extreme value 80.0, it made the mean and the median different and it doesn't affect the median but made the mean higher.

(c) ▷ In Urban home:

deleting in Urban sample

(the $x_{min} = 4.0$ and $x_{max} = 80.0$)

trimmed mean: $\bar{x}_u = \frac{237 - 80 - 4.0}{9} = 17.0$ (EU/mg)

corresponding trimming percentage:

$\frac{1}{11} \times 100\% \approx 9.1\%$

trimmed mean is less than the means without deleting the smallest and largest observation.

(sample is positive skew. without high extreme value, its mean will decrease)

The median doesn't change

▷ In Farm homes:

trimmed mean: $\bar{X}_F = \frac{128.4 - 0.3 - 21.0}{13} = \frac{107.1}{13} \approx 8.24$ (EU/mg)

trimming percentage: $\frac{1}{15} \times 100\% \approx 6.7\%$

trimmed mean is lower than mean of entrie sample, median is same value.

median, 25% trimmed mean, 10%
mple mean for the lifetime data given
mpare these measures.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

40. sample median: $\tilde{x} = \frac{91+93}{2} = 92$

25% trimmed mean: $\bar{x} = \frac{2377}{25} = 95.08$ (we need to throw the smallest number of 12.5 so I delete $\frac{65}{2}$ and $\frac{141}{2}$ and the remaining 12 numbers on both sides, then divide by 25.

10% trimmed mean: $\bar{x} = \frac{4089}{40} = 102.23$

sample mean: $\bar{x} = \frac{5963}{50} = 119.26$

Section 1.4

44. ...sumption During Fire ... Estimation" (*Ergonomics,* ... following data on oxygen consumption (mL/kg/min) for a sample of ten firefighters performing a fire-suppression simulation:

29.5 49.3 30.6 28.2 28.0 26.3 33.9 29.4 23.5 31.6

Compute the following:

a. The sample range
b. The sample variance s^2 from the definition (i.e., by first computing deviations, then squaring them, etc.)
c. The sample standard deviation
d. s^2 using the shortcut method

a) the sample range is $49.3 - 23.5 = 25.8$

b) since $\bar{x} = \frac{310.3}{10} = 31.03$

$x_1 - \bar{x} = -1.53$ $\quad x_6 - \bar{x} = -4.73$

$x_2 - \bar{x} = 18.27$ $\quad x_7 - \bar{x} = 2.87$

$x_3 - \bar{x} = -0.43$ $\quad x_8 - \bar{x} = -1.63$

$x_4 - \bar{x} = -2.83$ $\quad x_9 - \bar{x} = -7.53$

$x_5 - \bar{x} = -3.03$ $\quad x_{10} - \bar{x} = 0.57$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 443.801$$

$$s^2 = \frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{n-1} = \frac{443.801}{9} = 49.3112$$

(c) $s = \sqrt{s^2} \approx 7.0222$

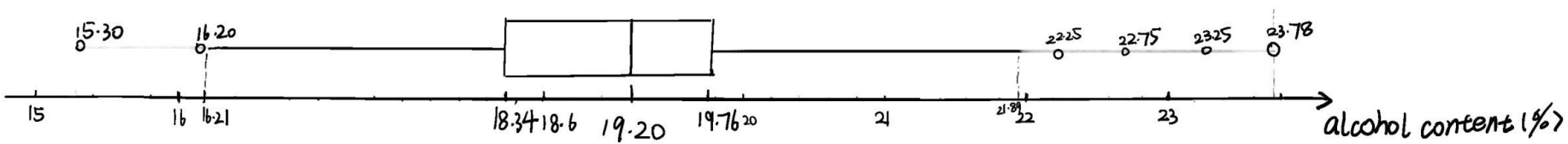
(d) $$s^2 = \frac{\sum x_i^2 - (\sum x_i)^2/n}{n-1}$$

$$= \frac{10072.41 - (310.3)^2/10}{9} = 49.3112$$

56. The following data on distilled alcohol content (%) for a sample of 35 port wines was extracted from the article "A Method for the Estimation of Alcohol in Fortified Wines Using Hydrometer Baumé and Refractometer Brix" (*Amer. J. Enol. Vitic.*, 2006: 486–490). Each value is an average of two duplicate measurements.

16.35 18.85 16.20 17.75 19.58 17.73 22.75 23.78 23.25
19.08 19.62 19.20 20.05 17.85 19.17 19.48 20.00 19.97
17.48 17.15 19.07 19.90 18.68 18.82 19.03 19.45 19.37
19.20 18.00 19.60 19.33 21.22 19.50 15.30 22.25

Use methods from this chapter, including a boxplot that shows outliers, to describe and summarize the data.



15.30 16.20 16.35 17.15 17.48 17.73 17.75 17.85 18.00 18.68 18.82 18.85 19.03 19.07 19.08 19.17 19.20 19.20 19.33 19.37 19.45 19.48 19.50 19.58 19.60 19.62 19.90
19.97 20.00 20.05 21.22 22.25 22.75 23.25 23.78

21.89

B+

median: 19.20

lower fourth: lower fourth is $(18.00 + 18.68) \times \frac{1}{2} = 18.34$

upper fourth: the upper fourth $(19.62 + 19.90) \times \frac{1}{2} = 19.76$.

$f_s = 19.76 - 18.34 = 1.42$ $1.5 f_s = 2.13$

upper $4^{th} + 1.5 f_s = 21.89$ upper $4^{th} + 3 f_s = 24.02$

lower $4^{th} - 1.5 f_s = 16.21$ lower $4^{th} - 3 f_s = 14.08$

that the outliers:

22.25 22.75 23.25 23.78

16.20 15.30

it show that some of the alcohol has been on the high side.