Week 1

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 : tens digit
leaf : ones digit

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

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)?

a)

2	3 2
3	4 7 3 9 5 4 6 7 2 8
4	6 0 8 3 8 1 5 9
5	1 1 6 8 0 4 0 4 5 0 6 1 6 5 9 7 0
6	7 9 4 2 6 4 5 3 2 0 9 6 1 0 7 2 6 4 6 9 8 9 2 0 3 0
7	1 0 5 5 6 3 5 5 6 2 2 4 3 0 5 8 0
8	0 8 3 2 4 4 3 2
9	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
18	9

stem : ones digit
leaf : tenth digit

b) tend to 7
c) highly concentrated
d) no, they are positively skewed
e) 18.9 is a outlier.

20. The article "Determination of Most Representative Subdivision" (*J. of Energy Engr.*, 1993: 43–55) gave data on various characteristics of subdivisions 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	2120	450	2250	2320	2400
3150	5700	5220	500	1850	2460	5850
2700	2730	1670	100	5770	3150	1890
510	240	396	1419	2109		

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?

a)

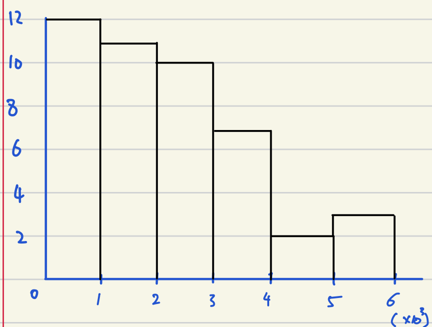
0	1 2 3 3 3 4 5 5 5 5 9 9
1	0 0 1 2 2 2 3 4 6 8 8
2	1 1 1 2 3 4 4 4 7 7
3	0 1 1 3 3 3 8
4	3 7
5	2 3 7 7 8

stem : thousands

leaf : hundreds

most data locate in the 2000's, and the graph is bimodal

b)



the shape is the same as the shape of the stem and leaf diagram

1.3

34. Exposure to microbial products, especially endotoxin, may have an impact on vulnerability to allergic diseases. The article "Dust Sampling Methods for Endotoxin—An Essential, But Underestimated Issue" (*Indoor Air,* 2006: 20–27) considered various issues associated 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.

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?

a) $\bar{U} = 21.55$

$\bar{F} = 8.56$

the mean of the urban sample is much higher than that of farm

b) U: 17.00

F: 8.90

because the data is urban area are concentrated in large numbers

c) $\bar{U}_{trimmed} = 17$

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

• the trimmed mean is less than the untrimmed mean

• the trimmed mean is the same as the median

$\bar{F}_{trimmed} = 8.24$

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

• trimmed mean is less than the original mean and median.

40. Compute the sample median, 25% trimmed mean, 10% trimmed mean, and sample mean for the lifetime data given in Exercise 27, and compare 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

median : 92

mean : 119.26

10% $\bar{x}_{trim}$: 102.23

25% $\bar{x}_{trim}$: 95.38

1.4

44. The article "Oxygen Consumption During Fire Suppression: Error of Heart Rate Estimation" (*Ergonomics,* 1991: 1469–1474) reported the 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) $49.3 - 23.5 = 25.8$

b) $\Sigma x_i = 310.3$

$\Sigma (x_i - \bar{x}) = 0$

$\Sigma (x_i - \bar{x})^2 = 443.801$

$\Sigma x_i^2 = 10072.41$

$\bar{x} = 31.03$

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

c) $\sqrt{49.3112} = 7.0222$

d) $$\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = 49.3112$$

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.

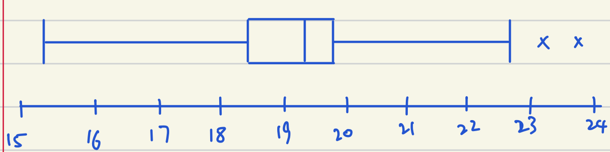
33

16.35	18.85	16.20	17.75	19.58	17.73	22.75	$\boxed{23.78}$	$\boxed{23.25}$
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 35
17 | 15 48 73 75 85
18 | 00 68 82 85
19 | 03 07 08 17 20 20 33 37 45 48 50 58 60 62 90 97
20 | 00 05
21 | 22
22 | 25 75

median : 17^{th} = 19.20
LQ : (18 + 18.68) ÷ 2 = 18.34
UQ : (19.9 + 19.62) ÷ 2 = 19.76
min : 15.30
max : 22.75



A