Solutions of suggested problems

CH2

The data in Table 2.13 are a sample from a larger data set collected on people discharged from a selected Pennsylvania hospital as part of a retrospective chart review of antibiotic usage in hospitals [7]. The data are also given in Data Set HOSPITAL.DAT with documentation in HOSPITAL.DOC at www.cengagebrain.com. Each data set at www.cengagebrain.com is available in six formats: ASCII, MINITAB-readable format, Excel-readable format, SAS-readable format, SPSS-readable format, and Stata-readable format, and as a text file (R-readable format).

2.1 Compute the mean and median for the duration of hospitalization for the 25 patients.

2.2 Compute the standard deviation and range for the duration of hospitalization for the 25 patients.

2.3 It is of clinical interest to know if the duration of hospitalization is affected by whether a patient has received antibiotics. Answer this question descriptively using either numeric or graphic methods.

Suppose the scale for a data set is changed by multiplying each observation by a positive constant.

***2.4** What is the effect on the median?

***2.5** What is the effect on the mode?

***2.6** What is the effect on the geometric mean?

***2.7** What is the effect on the range?

TABLE 2.13 Hospital-stay data

ID no.	Duration of hospital stay	Age	Sex 1 = M 2 = F	First temp. following admission	First WBC ($\times 10^3$) following admission	Received antibiotic? 1 = yes 2 = no	Received bacterial culture? 1 = yes 2 = no	Service 1 = med. 2 = surg.
1	5	30	2	99.0	8	2	2	1
2	10	73	2	98.0	5	2	1	1
3	6	40	2	99.0	12	2	2	2
4	11	47	2	98.2	4	2	2	2
5	5	25	2	98.5	11	2	2	2
6	14	82	1	96.8	6	1	2	2
7	30	60	1	99.5	8	1	1	1
8	11	56	2	98.6	7	2	2	1
9	17	43	2	98.0	7	2	2	1
10	3	50	1	98.0	12	2	1	2
11	9	59	2	97.6	7	2	1	1
12	3	4	1	97.8	3	2	2	2
13	8	22	2	99.5	11	1	2	2
14	8	33	2	98.4	14	1	1	2
15	5	20	2	98.4	11	2	1	2
16	5	32	1	99.0	9	2	2	2
17	7	36	1	99.2	6	1	2	2
18	4	69	1	98.0	6	2	2	2
19	3	47	1	97.0	5	1	2	1
20	7	22	1	98.2	6	2	2	2
21	9	11	1	98.2	10	2	2	2
22	11	19	1	98.6	14	1	2	2
23	11	67	2	97.6	4	2	2	1
24	9	43	2	98.6	5	2	2	2
25	4	41	2	98.0	5	2	2	1

2.1 We have

$$\bar{x} = \frac{\sum x_i}{n} = \frac{215}{25} = 8.6 \text{ days}$$

$$\text{median} = \frac{(n+1)}{2}\text{th largest observation} = 13\text{th largest observation} = 8 \text{ days}$$

2.2 We have that

$$s^2 = \frac{\sum_{i=1}^{25}(x_i - \bar{x})^2}{24} = \frac{(5-8.6)^2 + \cdots + (4-8.6)^2}{24} = \frac{784}{24} = 32.67$$

s = standard deviation = $\sqrt{\text{variance}}$ = 5.72 days
range = largest – smallest observation = 30 – 3 = 27 days

2.3 Suppose we divide the patients according to whether or not they received antibiotics, and calculate the mean and standard deviation for each of the two subsamples:

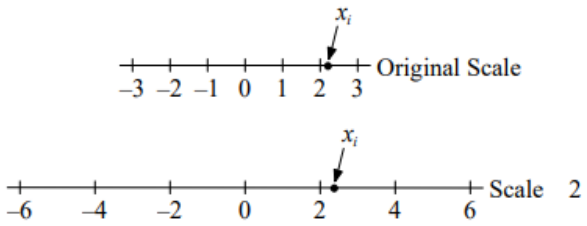
	$\bar{x}$	s	n
Antibiotics	11.57	8.81	7
No antibiotics	7.44	3.70	18
Antibiotics - x_7	8.50	3.73	6

It appears that antibiotic users stay longer in the hospital. Note that when we remove observation 7, the two standard deviations are in substantial agreement, and the difference in the means is not that impressive anymore. This example shows that $\bar{x}$ and s^2 are not robust; that is, their values are easily affected by outliers, particularly in small samples. Therefore, we would not conclude that hospital stay is different for antibiotic users vs. non-antibiotic users.

2.4-2.7 Changing the scale by a factor c will multiply each data value x_i by c, changing it to cx_i. Again the same individual's value will be at the median and the same individual's value will be at the mode, but these values will be multiplied by c. The geometric mean will be multiplied by c also, as can easily be shown:

$$\begin{aligned}\text{Geometric mean} &= [(cx_1)(cx_2)\cdots(cx_n)]^{1/n} \\ &= (c^n x_1 \cdot x_2 \cdots x_n)^{1/n} \\ &= c(x_1 \cdot x_2 \cdots x_n)^{1/n} \\ &= c \times \text{old geometric mean}\end{aligned}$$

The range will also be multiplied by c.
For example, if $c = 2$ we have:



A man runs 1 mile approximately once per weekend. He records his time over an 18-week period. The individual times and summary statistics are given in Table 2.14.

2.10 What is the mean and standard deviation of time_100?

TABLE 2.14 One mile running time for an individual, over 18 weeks

WK	Time (min)(x_i)	WK	Time (min)(x_i)
1	12.80	10	11.57
2	12.20	11	11.73
3	12.25	12	12.67
4	12.18	13	11.92
5	11.53	14	11.67
6	12.47	15	11.80
7	12.30	16	12.33
8	12.08	17	12.55
9	11.72	18	11.83

```
mean    1208.889

sd      38.74181
```

2.12 Suppose the man does not run for 6 months over the winter due to snow on the ground. He resumes running once a week in the spring and records a running time = 12.97 minutes in his first week of running in the spring.

2.12 The quantiles of the running times are

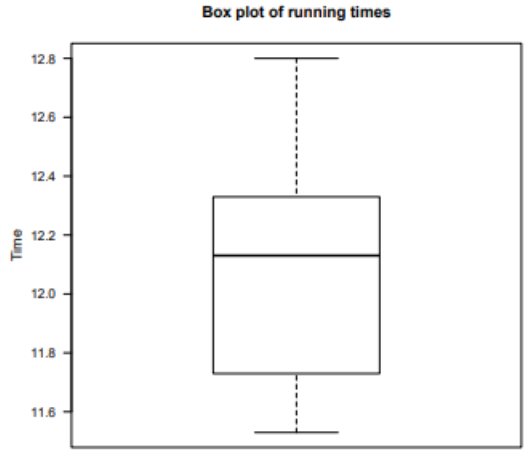
```
> quantile(running$time)
     0%      25%      50%      75%     100%
11.5300 11.7475 12.1300 12.3225 12.8000
```

An outlying value is identify has any value x such that
$x >$ upper quartile$+1.5\times$(upper quartile-lower quartile)

$=12.32+1.5\times(12.32-11.75)$

$=12.32+0.85=13.17$

Since 12.97 minutes is smaller than the largest nonoutlying value (13.17 minutes), this running time recorded in his first week of running in the spring is not an outlying value relative to the distribution of running times recorded the previous year.



The data in Table 2.15 are a sample of cholesterol levels taken from 24 hospital employees who were on a standard American diet and who agreed to adopt a vegetarian diet for 1 month. Serum-cholesterol measurements were made before adopting the diet and 1 month after. The data are available at cholesterol.xls at www.cengagebrain.com.

***2.13** Compute the mean change in cholesterol.

***2.14** Compute the standard deviation of the change in cholesterol levels.

2.15 Construct a stem-and-leaf plot of the cholesterol changes.

***2.16** Compute the median change in cholesterol.

2.17 Construct a box plot of the cholesterol changes to the right of the stem-and-leaf plot.

2.18 Some investigators believe that the effects of diet on cholesterol are more evident in people with high rather than low cholesterol levels. If you split the data in Table 2.15 according to whether baseline cholesterol is above or below the median, can you comment descriptively on this issue?

TABLE 2.15 Serum-cholesterol levels (mg/dL) before and after adopting a vegetarian diet

Subject	Before	After	Difference*
1	195	146	49
2	145	155	−10
3	205	178	27
4	159	146	13
5	244	208	36
6	166	147	19
7	250	202	48
8	236	215	21
9	192	184	8
10	224	208	16
11	238	206	32
12	197	169	28
13	169	182	−13
14	158	127	31
15	151	149	2
16	197	178	19
17	180	161	19
18	222	187	35
19	168	176	−8
20	168	145	23
21	167	154	13
22	161	153	8
23	178	137	41
24	137	125	12

*Before – after.

2.13 The mean is

$$\bar{x} = \frac{\sum x_i}{24} = \frac{469}{24} = 19.54 \text{ mg/dL}$$

2.14 We have that

$$s^2 = \frac{\sum_{i=1}^{24}(x_i - \bar{x})^2}{23} = \frac{(49-19.54)^2 + \cdots + (12-19.54)^2}{23} = \frac{6495.96}{23} = 282.43$$

$$s = \sqrt{282.43} = 16.81 \text{ mg/dL}$$

2.15 We provide two rows for each stem corresponding to leaves 5-9 and 0-4 respectively. We have

Stem-and-leaf plot		Cumulative frequency
+4	98	24
+4	1	22
+3	65	21
+3	21	19
+2	78	17
+2	13	15
+1	9699	13
+1	332	9
+0	88	6
+0	2	4
−0		
−0	8	3
−1	03	2

2.16 We wish to compute the average of the (24/2)th and (24/2 + 1)th largest values = average of the 12th and 13th largest points. We note from the stem-and-leaf plot that the 13th largest point counting from the bottom is the largest value in the upper +1 row = 19. The 12th largest point = the next largest value in this row = 19. Thus, the median $= \frac{19+19}{2} = 19$ mg/dL.

2.17 We first must compute the upper and lower quartiles. Because $24(75/100) = 18$ is an integer, the upper quartile = average of the 18th and 19th largest values $= \frac{32+31}{2} = 31.5$. Similarly, because $24(25/100) = 6$ is an integer, the lower quartile = average of the 6th and 7th smallest points $= \frac{8+12}{2} = 10$.

Second, we identify outlying values. An outlying value is identified as any value x such that

$$\begin{aligned} x &> \text{upper quartile} + 1.5 \times (\text{upper quartile} - \text{lower quartile}) \\ &= 31.5 + 1.5 \times (31.5 - 10) \\ &= 31.5 + 32.25 = 63.75 \end{aligned}$$

or

$$\begin{aligned} x &< \text{lower quartile} - 1.5 \times (\text{upper quartile} - \text{lower quartile}) \\ &= 10 - 1.5 \times (31.5 - 10) \\ &= 10 - 32.25 = -22.25 \end{aligned}$$

From the stem-and-leaf plot, we note that the range is from −13 to +49. Therefore, there are no outlying values. Thus, the box plot is as follows:

Stem-and-leaf plot		Cumulative frequency	Box plot
+4	98	24	|
+4	1	22	|
+3	65	21	|
+3	21	19	+−−−−−+
+2	78	17	| |
+2	13	15	| |
+1	9699	13	*−−+−−*
+1	332	9	+−−−−−+
+0	88	6	|
+0	2	4	|
−0			|
−0	8	3	|
−1	03	2	|

Comments: The distribution is reasonably symmetric, since the mean = 19.54 mg/dL ≐ 19 mg/dL = median. This is also manifested by the percentiles of the distribution since the upper quartile −median = 31.5 − 19 = 12.5 ≐ median − lower quartile = 19 − 10 = 9. The box plot looks deceptively asymmetric, since 19 is the highest value in the upper +1 row and 10 is the lowest value in the lower +1 row.

2.18 To compute the median cholesterol level, we construct a stem-and-leaf plot of the before-cholesterol measurements as follows.

Stem-and-leaf plot		Cumulative frequency
25	0	24
24	4	23
23	68	22
22	42	20
21		
20	5	18
19	5277	17
18	0	13
17	8	12
16	698871	11
15	981	5
14	5	2
13	7	1

Based on the cumulative frequency column, we see that the median = average of the 12th and 13th largest values $= \frac{178+180}{2} = 179$ mg/dL. Therefore, we look at the change scores among persons with baseline cholesterol ≥ 179 mg/dL and < 179 mg/dL, respectively. A stem-and-leaf plot of the change scores in these two groups is given as follows:

Baseline ≥ 179 mg/dL		Baseline < 179 mg/dL	
Stem-and-leaf plot		Stem-and-leaf plot	
+4	98	+4	
+4		+4	1
+3	65	+3	
+3	2	+3	1
+2	78	+2	
+2	1	+2	3
+1	699	+1	9
+1		+1	332
+0	8	+0	8
+0		+0	2
−0		−0	
−0		−0	8
−1		−1	03

Clearly, from the plot, the effect of diet on cholesterol is much greater among individuals who start with relatively high cholesterol levels (≥ 179 mg/dL) versus those who start with relatively low levels (< 179 mg/dL). This is also evidenced by the mean change in cholesterol levels in the two groups, which is 28.2 mg/dL in the ≥ 179 mg/dL group and 10.9 mg/dL in the < 179 mg/dL group. We will be discussing the formal statistical methods for comparing mean changes in two groups in our work on two-sample inference in Chapter 8.