Complete Solutions Manual to Accompany

Fundamentals of Biostatistics

EIGHTH EDITION

Bernard Rosner
Harvard University,
Cambridge, MA

Prepared by

Roland A. Matsouaka
Duke University, Durham, NC

CENGAGE Learning

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

CENGAGE Learning

For product information and technology assistance, contact us at **Cengage Learning Customer & Sales Support, 1-800-354-9706**.

For permission to use material from this text or product, submit all requests online at **www.cengage.com/permissions**
Further permissions questions can be emailed to **permissionrequest@cengage.com**.

ISBN-13: 978-1-305-26905-7
ISBN-10: 1-305-26905-5

Cengage Learning
20 Channel Center Street, 4th Floor
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at: **www.cengage.com/global**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**.

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**.

Printed in the United States of America
1 2 3 4 5 6 7 17 16 15 14 13

Contents

Chapter 2

DESCRIPTIVE STATISTICS

2.1 We have

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

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

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:

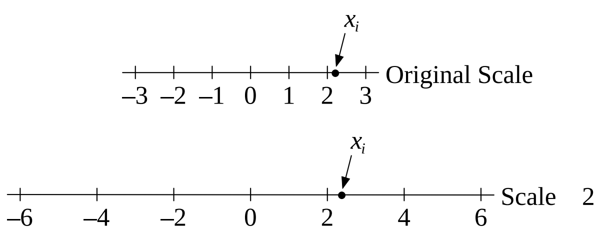
	$\bar{x}$	***s***	***n***
Antibiotics	11.57	8.81	7
No antibiotics	7.44	3.70	18
Antibiotics - x_7	8.50	3.73	6

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:



2.8 We first read the data file "running time" in R

```
> require(xlsx)
> running<-na.omit(read.xlsx("C:/Data_sets/running_time.xlsx",1,
  header=TRUE))
```

Let us print the first observations

```
> head(running)
  week  time
1    1 12.80
2    2 12.20
3    3 12.25
4    4 12.18
5    5 11.53
6    6 12.47
```

The mean 1-mile running time over 18 weeks is equal to 12.09 minutes:

```
> mean(running$time)
[1] 12.08889
```

2.9 The standard deviation is given by

```
> sd(running$time)
[1] 0.3874181
```

2.10 Let us first create the variable "time_100" and then calculate its mean and standard deviation

```
> running$time_100=100*running$time
> mean(running$time_100)
[1] 1208.889

> sd(running$time_100)
[1] 38.74181
```

2.11 Let us to construct the stem-and-leaf plot in R using the stem.leaf command from the package "aplpack"

```
> require(aplpack)
```

```
> stem.leaf(running$time_100, unit=1, trim.outliers=FALSE)

1 | 2: represents 12
 leaf unit: 1
            n: 18
   2    115 | 37
   3    116 | 7
   5    117 | 23
   7    118 | 03
   8    119 | 2
  (1)   120 | 8
   9    121 | 8
   8    122 | 05
   6    123 | 03
   4    124 | 7
   3    125 | 5
   2    126 | 7
        127 |
   1    128 | 0
```

Note: one can also use the standard command stem (which does require the "aplpack" package) to get a similar plot

```
> stem(running$time_100, scale = 4)
```

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
```

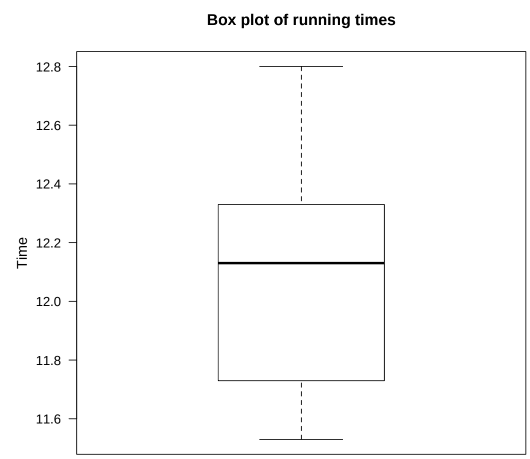
An outlying value is identify has any value x such that

$x >$ 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.



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 $= \frac{19+19}{2} = 19\ \text{mg/dL}$.

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:

<table>
<tr><th colspan="2">Stem-and-leaf plot</th><th>Cumulative frequency</th><th>Box plot</th></tr>
<tr><td>+4</td><td>98</td><td>24</td><td>|</td></tr>
<tr><td>+4</td><td>1</td><td>22</td><td>|</td></tr>
<tr><td>+3</td><td>65</td><td>21</td><td>|</td></tr>
<tr><td>+3</td><td>21</td><td>19</td><td>+ – – – – – +</td></tr>
<tr><td>+2</td><td>78</td><td>17</td><td>| |</td></tr>
<tr><td>+2</td><td>13</td><td>15</td><td>| |</td></tr>
<tr><td>+1</td><td>9699</td><td>13</td><td>* – – + – – *</td></tr>
<tr><td>+1</td><td>332</td><td>9</td><td>+ – – – – – +</td></tr>
<tr><td>+0</td><td>88</td><td>6</td><td>|</td></tr>
<tr><td>+0</td><td>2</td><td>4</td><td>|</td></tr>
<tr><td>–0</td><td></td><td></td><td>|</td></tr>
<tr><td>–0</td><td>8</td><td>3</td><td>|</td></tr>
<tr><td>–1</td><td>03</td><td>2</td><td>|</td></tr>
</table>

Comments: The distribution is reasonably symmetric, since the mean $= 19.54$ mg/dL $\doteq 19$ mg/dL = median. This is also manifested by the percentiles of the distribution since the upper quartile $-$median $= 31.5 - 19 = 12.5 \doteq$ 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 Stem-and-leaf plot		Baseline < 179 mg/dL 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 ($\geq$ 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 $\geq$ 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.

2.19 We first calculate the difference scores between the two positions:

Subject number	Subject	Systolic difference score	Diastolic difference score
1	B.R.A.	–6	–8
2	J.A.B.	+2	–2
3	F.L.B.	+6	+4
4	V.P.B.	+8	–4
5	M.F.B.	+8	+2
6	E.H.B.	+12	+4
7	G.C.	+10	0
8	M.M.C.	0	–2
9	T.J.F.	–2	–8
10	R.R.F.	+4	–2
11	C.R.F.	+8	–2
12	E.W.G.	+14	+4
13	T.F.H.	+2	–14
14	E.J.H.	+6	–2
15	H.B.H.	+26	0
16	R.T.K.	+8	+8
17	W.E.L.	+10	+4
18	R.L.L.	+12	+2
19	H.S.M.	+14	+8
20	V.J.M.	–8	–2
21	R.H.P.	+10	+14
22	R.C.R.	+14	+4
23	J.A.R.	+14	0
24	A.K.R.	+4	+4
25	T.H.S.	+6	+4
26	O.E.S.	+16	+2
27	R.E.S.	+28	+16
28	E.C.T.	+18	–4
29	J.H.T.	+14	+4
30	F.P.V.	+4	–6
31	P.F.W.	+12	+6
32	W.J.W.	+8	–4

Second, we calculate the mean difference scores:

$$\bar{x}_{\text{sys}} = \frac{-6+\ldots+8}{32} = \frac{282}{32} = 8.8 \text{ mm Hg}$$

$$\bar{x}_{\text{dias}} = \frac{-8+\ldots+(-4)}{32} = \frac{30}{32} = 0.9 \text{ mm Hg}$$

The median difference scores are given by the average of the 16th and 17th largest values. Thus,

$$\text{median}_{\text{sys}} = \frac{8+8}{2} = 8 \text{ mm Hg}$$

$$\text{median}_{\text{dias}} = \frac{0+2}{2} = 1 \text{ mm Hg}$$

2.20 The stem-and-leaf and box plots allowing two rows for each stem are given as follows:

Systolic Blood Pressure

	Stem-and-leaf plot	Cumulative frequency	Box plot
2	68	32	|
2			|
1	68	30	|
1	20402404442	28	+ − − − − − +
0	68886868	17	* − − + − − *
0	204244	9	+ − − − − − +
–0	2	3	|
–0	68	2	|

Median $= 8$, upper quartile $= \dfrac{14+14}{2} = 14$, lower quartile $= \dfrac{4+4}{2} = 4$, outlying values: $x > 14 + 1.5 \times (14 - 4) = 29$ or $x < 4 - 1.5 \times (14 - 4) = -11$. Since the range of values is from –8 to +28, there are no outlying values for systolic blood pressure.

Diastolic Blood Pressure

	Stem-and-leaf plot	Cumulative frequency	Box plot
1	6	32	0
1	4	31	0
0	886	30	|
0	42404042404424	27	+ − − + − − +
–0	242222244	13	+ − − − − − +
–0	886	4	|
–1	4	1	0

Median $= 1$, upper quartile $= \dfrac{4+4}{2} = 4$, lower quartile $= \dfrac{-2-2}{2} = -2$, outlying values: $x > 4 + 1.5 \times (4+2) = 13.0$ or $x < -2 - 1.5 \times (4+2) = -11.0$. The values +16, +14 and –14 are outlying values.

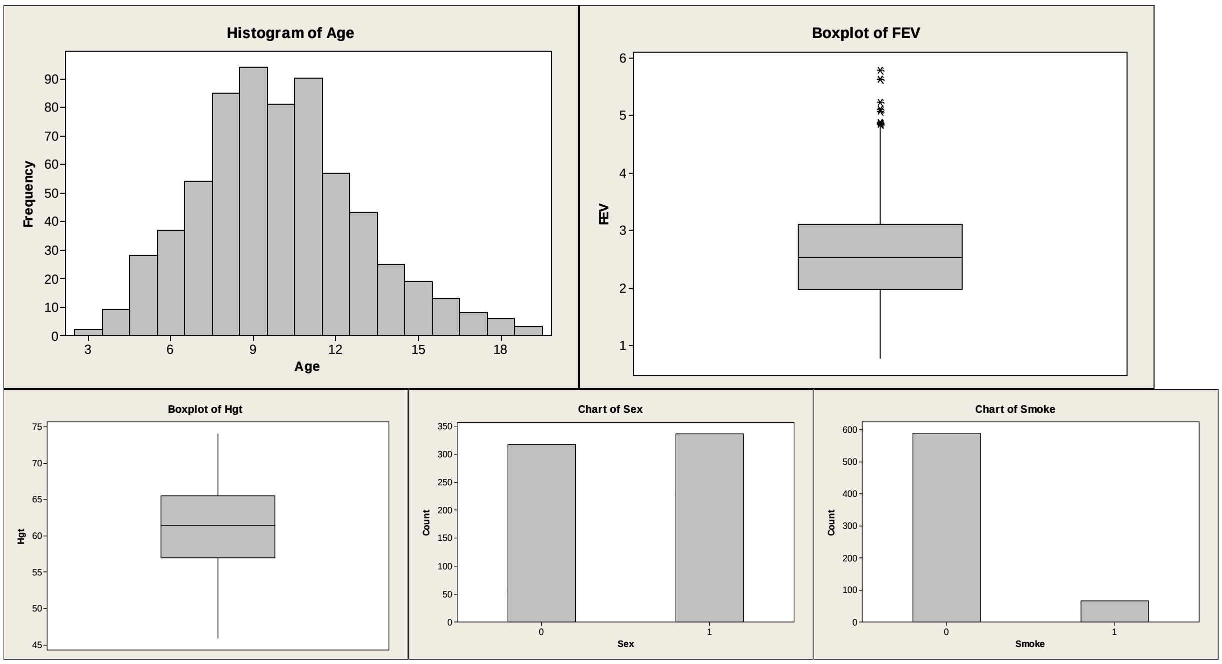
2.21 Systolic blood pressure clearly seems to be higher in the supine (recumbent) position than in the standing position. Diastolic blood pressure appears to be comparable in the two positions. The distributions are each reasonably symmetric.

2.22 The upper and lower deciles for postural change in systolic blood pressure (SBP) are 14 and 0. Thus, the normal range for postural change in SBP is $0 \le x \le 14$. The upper and lower deciles for postural change in diastolic blood pressure (DBP) are 8 and –6. Thus, the normal range for postural change in DBP is $-6 \le x \le 8$.

2.23

```
   Id       Age       FEV       Hgt       Sex     Smoke
  301         9     1.708        57         0         0
  451         8     1.724      67.5         0         0
......
61951        15     2.278        60         0         1
63241        16     4.504        72         1         0
71141        17     5.638        70         1         0
71142        16     4.872        72         1         1
73041        16      4.27        67         1         1
73042        15     3.727        68         1         1
73751        18     2.853        60         0         0
```

75852	16	2.795	63	0	1
77151	15	3.211	66.5	0	0
MEAN	9.931193	2.63678	61.14358	0.513761	0.099388
MEDIAN	10	2.5475	61.5		
SD	2.953935	0.867059	5.703513		



2.24 Results for Sex = 0

Variable	Age	Mean	StDev	Minimum	Median	Maximum
FEV	3	1.0720	*	1.0720	1.0720	1.0720
	4	1.316	0.290	0.839	1.404	1.577
	5	1.3599	0.2513	0.7910	1.3715	1.7040
	6	1.6477	0.2182	1.3380	1.6720	2.1020
	7	1.8330	0.3136	1.3700	1.7420	2.5640
	8	2.1490	0.4046	1.2920	2.1900	2.9930
	9	2.3753	0.4407	1.5910	2.3810	3.2230
	10	2.6814	0.4304	1.4580	2.6895	3.4130
	11	2.8482	0.4293	2.0810	2.8220	3.7740
	12	2.9481	0.3679	2.3470	2.8890	3.8350
	13	3.0656	0.4321	2.2160	3.1135	3.8160
	14	2.962	0.383	2.236	2.997	3.428
	15	2.761	0.415	2.198	2.783	3.330
	16	3.058	0.397	2.608	2.942	3.674
	17	3.5000	*	3.5000	3.5000	3.5000
	18	2.9470	0.1199	2.8530	2.9060	3.0820
	19	3.4320	0.1230	3.3450	3.4320	3.5190

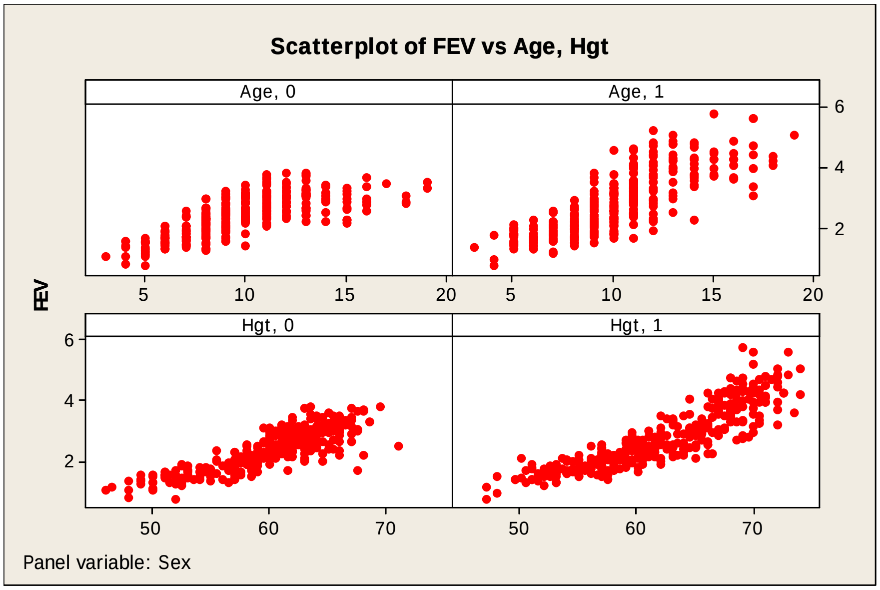
Results for Sex = 1

Variable	Age	Mean	StDev	Minimum	Median	Maximum
FEV	3	1.4040	*	1.4040	1.4040	1.4040
	4	1.196	0.524	0.796	1.004	1.789
	5	1.7447	0.2336	1.3590	1.7920	2.1150
	6	1.6650	0.2304	1.3380	1.6580	2.2620

7	1.9117	0.3594	1.1650	1.9050	2.5780
8	2.0756	0.3767	1.4290	2.0690	2.9270
9	2.4822	0.5086	1.5580	2.4570	3.8420
10	2.6965	0.6020	1.6650	2.6080	4.5910
11	3.2304	0.6459	1.6940	3.2060	4.6370
12	3.509	0.871	1.916	3.530	5.224
13	4.011	0.690	2.531	4.045	5.083
14	3.931	0.635	2.276	3.882	4.842
15	4.289	0.644	3.727	4.279	5.793
16	4.193	0.437	3.645	4.270	4.872
17	4.410	1.006	3.082	4.429	5.638
18	4.2367	0.1597	4.0860	4.2200	4.4040
19	5.1020	*	5.1020	5.1020	5.1020

Results for Sex = 0

Variable	Hgt	Mean
FEV	46.0	1.0720
	46.5	1.1960
	48.0	1.110
	49.0	1.4193
	50.0	1.3378
	51.0	1.5800
	51.5	1.474
	52.0	1.389
	52.5	1.577
	53.0	1.6887
	53.5	1.4150
	54.0	1.6408
	54.5	1.7483
	55.0	1.6313
	55.5	2.036
	56.0	1.651
	56.5	1.7875
	57.0	1.9037
	57.5	1.9300
	58.0	2.1934
	58.5	1.9440
	59.0	2.1996
	59.5	2.517
	60.0	2.5659
	60.5	2.5563
	61.0	2.6981
	61.5	2.626
	62.0	2.7861
	62.5	2.7777
	63.0	2.7266
	63.5	2.995
	64.0	2.9731
	64.5	2.864
	65.0	3.090
	65.4	2.4340
	65.5	3.154
	66.0	2.984
	66.5	3.2843
	67.0	3.167
	67.5	2.922
	68.0	3.214
	68.5	3.3300
	69.5	3.8350
	71.0	2.5380



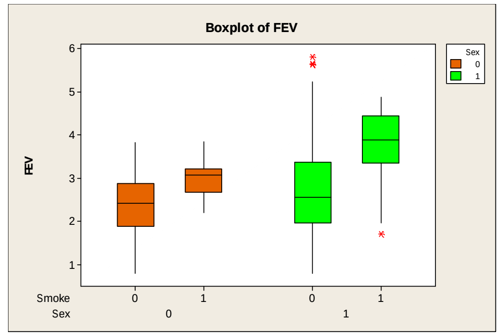
Results for Sex = 1

Variable	Hgt	Mean
FEV	47.0	0.981
	48.0	1.270
	49.5	1.4250
	50.0	1.794
	50.5	1.536
	51.0	1.683
	51.5	1.514
	52.0	1.5915
	52.5	1.7100
	53.0	1.6646
	53.5	1.974
	54.0	1.7809
	54.5	1.8380
	55.0	1.8034
	55.5	1.8070
	56.0	2.025
	56.5	1.879
	57.0	2.0875
	57.5	1.829
	58.0	2.0169
	58.5	2.131
	59.0	2.350
	59.5	2.515
	60.0	2.279
	60.5	2.3253
	61.0	2.4699
	61.5	2.5410
	62.0	2.658
	62.5	2.829
	63.0	2.877
	63.5	2.757
	64.0	2.697
	64.5	3.100
	65.0	2.770
	65.5	3.0343
	66.0	3.115
	66.5	3.353
	67.0	3.779
	67.5	3.612
	68.0	3.878
	68.5	3.872
	69.0	4.022
	69.5	3.743
	70.0	4.197
	70.5	3.931
	71.0	4.310
	71.5	4.7200
	72.0	4.361
	72.5	4.2720
	73.0	5.255
	73.5	3.6450
	74.0	4.654

Descriptive Statistics: FEV

Results for Sex = 0

Variable	Smoke	Mean	StDev
FEV	0	2.3792	0.6393
	1	2.9659	0.4229



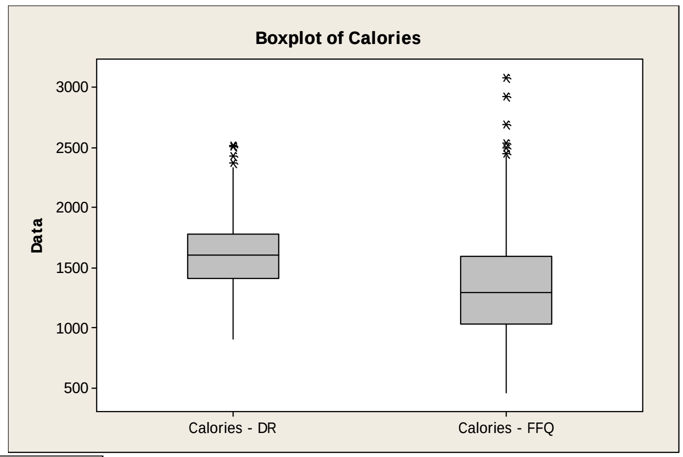
Results for Sex = 1

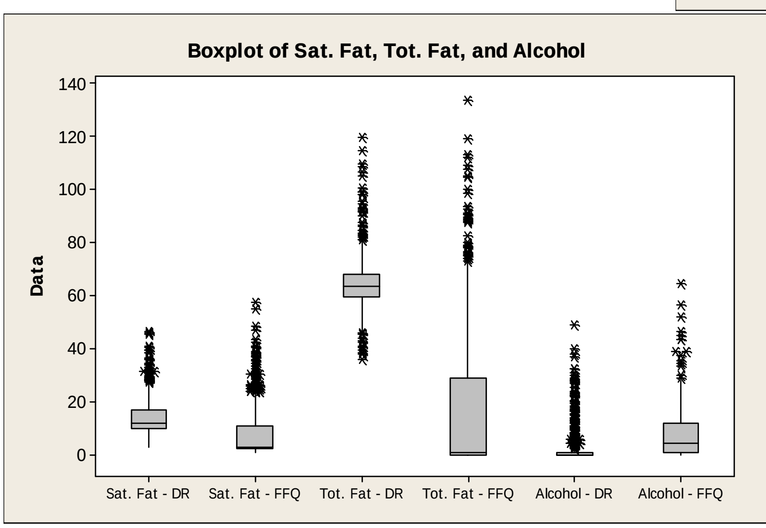
Variable	Smoke	Mean	StDev
FEV	0	2.7344	0.9741
	1	3.743	0.889

2.25 Looking at the scatterplot of FEV vs. Age, we find that FEV increases with age for both boys and girls, at approximately the same rate. However, the spread (standard deviation) of FEV values appears to be higher in male group than in the female group.

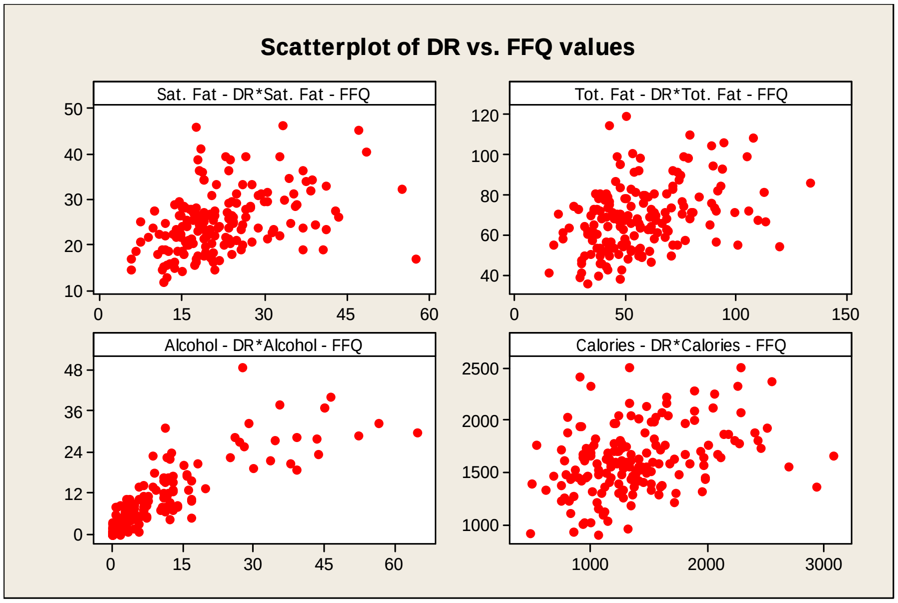
2.26

Variable	Mean	StDev	Median
Sat. Fat - DR	14.557	7.536	12.000
Sat. Fat - FFQ	7.898	9.695	3.159
Tot. Fat - DR	64.238	9.894	63.500
Tot. Fat - FFQ	15.21	27.00	1.00
Alcohol - DR	2.470	6.314	0.000
Alcohol - FFQ	8.951	12.255	4.550
Calories - DR	1619.9	323.4	1606.0
Calories - FFQ	1371.7	482.1	1297.6





2.27



If FFQ were a perfect substitute for DR, the points would line up in a straight line. If the two were unrelated, then we would expect to see a random pattern in each panel. The scatterplots shown above seem to suggest that the DR and FFQ values are not highly related.

2.28 The 5x5 tables below show the number of people classified into a particular combination of quintile categories. For each table, the rows represent the quintiles of the DR, and the columns represent quintiles of the FFQ. Overall, we get the same impression that there is weak concordance between the two measures. However, we do notice that the agreement is greatest for the two measures with regards to alcohol consumption. Also, we note the relatively high level of agreement at the extremes of each nutrient; for example, the (1,1) and (5,5) cells generally contain the highest values.

Tabulated statistics: SFDQuin, SFFQuin

Rows: SFDQuin   Columns: SFFQuin

	1	2	3	4	5	All
1	15	8	9	2	1	35
2	10	6	6	8	5	35
3	4	7	8	9	6	34
4	6	10	6	9	4	35
5	0	3	6	7	18	34
All	35	34	35	35	34	173

Cell Contents: Count

Tabulated statistics: TFDQuin, TFFQuin

Rows: TFDQuin   Columns: TFFQuin

	1	2	3	4	5	All
1	13	9	8	5	1	36
2	9	5	7	10	3	34
3	4	10	8	6	6	34
4	8	6	3	9	9	35
5	1	5	8	5	15	34

All	35	35	34	35	34	173

Cell Contents: Count

Tabulated statistics: AlcDQuin, AlcFQuin

Rows: AlcDQuin Columns: AlcFQuin

	1	2	3	4	5	All
1	28	5	2	0	0	35
2	6	23	6	0	0	35
3	0	9	14	10	1	34
4	0	1	10	16	8	35
5	0	0	0	8	26	34
All	34	38	32	34	35	173

Cell Contents: Count

Tabulated statistics: CalDQuin, CalFQuin

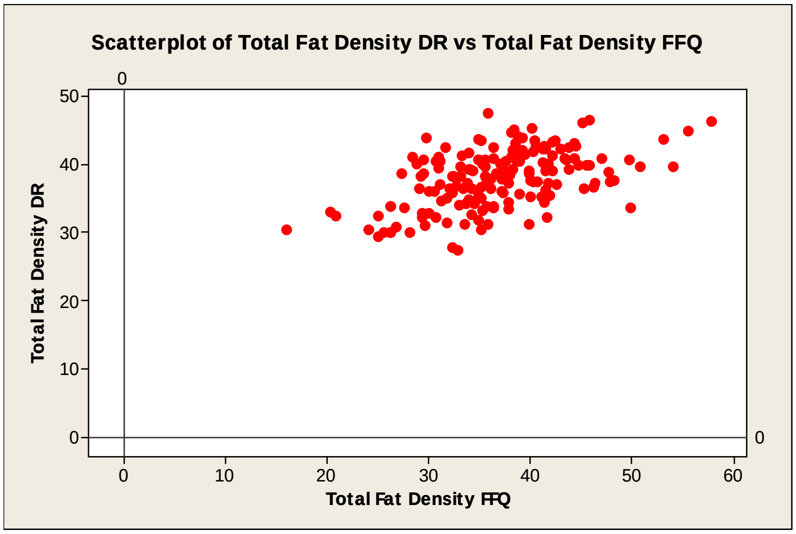
Rows: CalDQuin Columns: CalFQuin

	1	2	3	4	5	All
1	10	11	8	4	2	35
2	11	4	9	7	4	35
3	5	9	6	8	6	34
4	4	8	7	6	10	35
5	5	3	4	10	12	34
All	35	35	34	35	34	173

2.29

Descriptive Statistics: Total Fat Density DR, Total Fat Density FFQ

Variable	Mean	StDev	Median
Total Fat Density DR	38.066	4.205	38.646
Total Fat Density FFQ	36.855	6.729	36.366



2.30 The concordance for the quintiles of nutrient density does appear somewhat stronger than for the quintiles of raw nutrient data. In the table below, we see that 19+14+10+7+11 = 61 individuals were in the same quintile on both measures, compared to 50 people in the table from question 2.28.

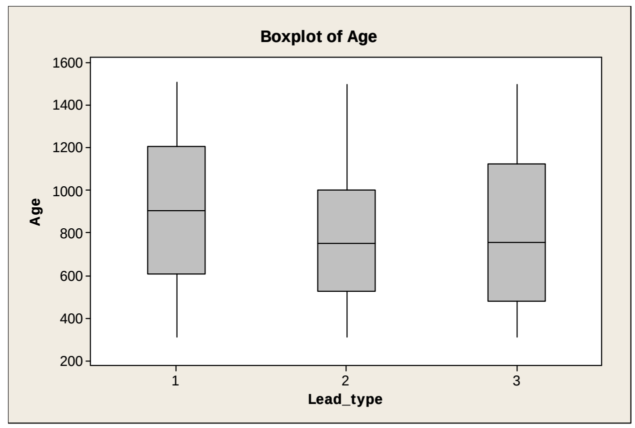
Tabulated statistics: Dens DR Quin, Dens FFQ Quin

Rows: Dens DR Quin Columns: Dens FFQ Quin

	1	2	3	4	5	All
1	19	7	6	2	1	35
2	5	14	5	6	5	35
3	4	8	10	6	6	34
4	6	4	7	7	11	35
5	1	2	6	14	11	34
All	35	35	34	35	34	173

2.31 We find that exposed children (Lead type = 2) are somewhat younger and more likely to be male (Sex = 1), compared to unexposed children. The boxplot below shows all three lead types, but we are only interested in types 1 and 2.

Variable	Lead_type	Mean	StDev	Median
Age	1	893.8	360.2	905.0
	2	776.3	329.5	753.5



Tabulated statistics: Lead_type, Sex

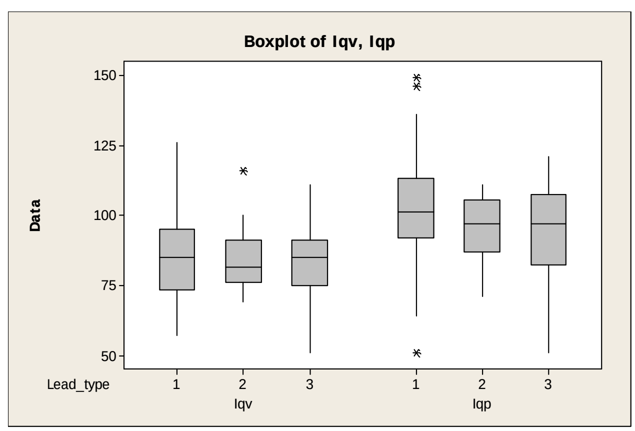
Rows: Lead_type Columns: Sex

	1	2	All
1	46	32	78
	58.97	41.03	100.00
2	17	7	24
	70.83	29.17	100.00

2.32 The exposed children have somewhat lower mean and median IQ scores compared to the unexposed children, but the differences don't appear to be very large.

Descriptive Statistics: Iqv, Iqp

Variable	Lead_type	Mean	StDev	Median
Iqv	1	85.14	14.69	85.00
	2	84.33	10.55	81.50
Iqp	1	102.71	16.79	101.00
	2	95.67	11.34	97.00



2.33 The coefficient of variation (CV) is given by 100% $(s/\bar{x})$, where s and $\bar{x}$ are computed separately for each subject. We compute $\bar{x}$, s, and $\mathrm{CV} = 100\% \times (s/\bar{x})$ separately for each subject using the following function in R:

```
 cv_est<-function(x){
  m=mean(x)
  s=sd(x)
  cv=100*(s/m)
 cat("The mean, SD, CV are \n")
 return(c(m, s, cv))
}
```

For the first subject, we have

```
> cv_est(c(2.22, 1.88))
Mean, SD, CV are
[1]  2.0500000  0.2404163 11.7276247
```

The results are shown in the table below:

APC resistance Coefficient of Variation

Sample number	A	B	mean	sd	CV
1	2.22	1.88	2.05	0.240	11.7
2	3.42	3.59	3.505	0.120	3.4
3	3.68	3.01	3.345	0.474	14.2
4	2.64	2.37	2.505	0.191	7.6
5	2.68	2.26	2.47	0.297	12.0
6	3.29	3.04	3.165	0.177	5.6
7	3.85	3.57	3.71	0.198	5.3
8	2.24	2.29	2.265	0.035	1.6
9	3.25	3.39	3.32	0.099	3.0
10	3.3	3.16	3.23	0.099	3.1
				average CV	6.7

2.34 To obtain the average CV, we average the individual-specific CV's over the 10. The average CV = 6.7% which indicates excellent reproducibility.

2.35 We compute the mean and standard deviation of pod weight for both inoculated (I) and uninoculated (U) plants. The results are given as follows:

	I	**U**
mean	1.63	1.08
sd	0.42	0.51
n	8	8

2.36 We plot the distribution of I and U pod weights using a dot-plot from MINITAB.

```
                .     .     .  .     ..    .          .
-------+---------+---------+---------+---------+---------- I

                .
 .      .  . .  .              .               .
-------+---------+---------+---------+---------+---------- U
     0.70      1.05      1.40      1.75      2.10      2.45
```

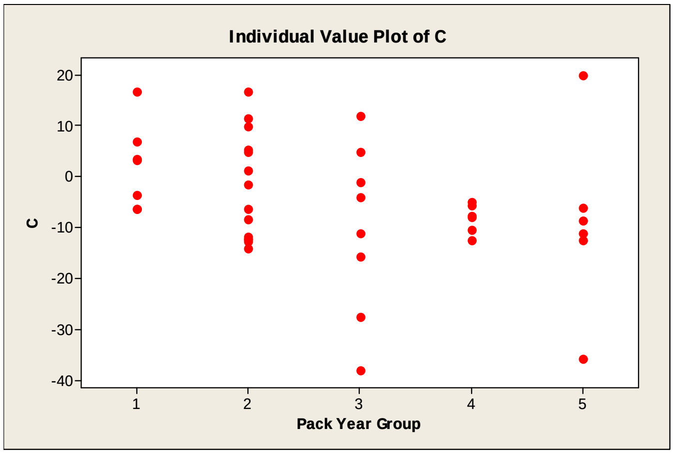
2.37 Although there is some overlap in the distributions, it appears that the I plants tend in have higher pod weights than the U plants. We will discuss *t* tests in Chapter 8 to assess whether there are "statistically significant" differences in mean pod weights between the 2 groups.

2.38-2.40 For lumbar spine bone mineral density, we have the following:

ID	A	B	C	PY Diff	Pack Year Group
1002501	-0.05	0.785	-6.36942675	13.75	2
1015401	-0.12	0.95	-12.6315789	48	5
1027601	-0.24	0.63	-38.0952381	20.5	3
1034301	0.04	0.83	4.81927711	29.75	3
1121202	-0.19	0.685	-27.7372263	25	3
1162502	-0.03	0.845	-3.55029586	5	1
1188701	-0.08	0.91	-8.79120879	42	5
1248202	-0.1	0.71	-14.084507	15	2
1268301	0.15	0.905	16.5745856	9.5	1
1269402	-0.12	0.95	-12.6315789	39	4
1273101	-0.1	0.81	-12.345679	14.5	2
1323501	0.09	0.755	11.9205298	23.25	3
1337102	-0.08	0.67	-11.9402985	18.5	2
1467301	-0.07	0.665	-10.5263158	39	4
1479401	-0.03	0.715	-4.1958042	25.5	3
1494101	0.05	0.735	6.80272109	8	1
1497701	0.04	0.75	5.33333333	10	2
1505502	-0.04	0.81	-4.9382716	32	4
1519402	-0.01	0.645	-1.5503876	13.2	2
1521701	-0.06	0.74	-8.10810811	30	4
1528201	-0.11	0.695	-15.8273381	20.25	3
1536201	-0.05	0.865	-5.78034682	36.25	4
1536701	0.03	0.635	4.72440945	12	2
1541902	-0.12	0.98	-12.244898	11.25	2
1543602	0.03	0.885	3.38983051	8	1
1596702	0.01	0.955	1.04712042	14	2
1597002	0.07	0.705	9.92907801	17.3	2
1597601	0.13	0.775	16.7741935	12	2
1607901	-0.03	0.485	-6.18556701	43.2	5
1608801	-0.21	0.585	-35.8974359	48	5
1628601	-0.05	0.795	-6.28930818	5.35	1
1635901	0.03	0.945	3.17460317	8	1
1637901	-0.05	0.775	-6.4516129	6	1
1640701	-0.01	0.855	-1.16959064	28	3
1643602	0.11	0.555	19.8198198	64.5	5
1647502	-0.07	0.545	-12.8440367	11.3	2
1648701	-0.08	0.94	-8.5106383	15.75	2
1657301	-0.08	0.72	-11.1111111	21	3
1671001	-0.07	0.895	-7.82122905	39	4
1672702	0.1	0.87	11.4942529	18.75	2
2609801	-0.1	0.9	-11.1111111	48	5
		Mean	-4.9496682		
		Median	-6.2893082		
		Sd	12.4834202		

Descriptive Statistics: C

Variable	Pack Year Group	Mean	StDev	Median
C	1	1.95	8.26	3.17
	2	-2.18	10.45	-3.96
	3	-10.17	16.69	-7.65
	4	-8.30	2.89	-7.96
	5	-9.13	17.77	-9.95



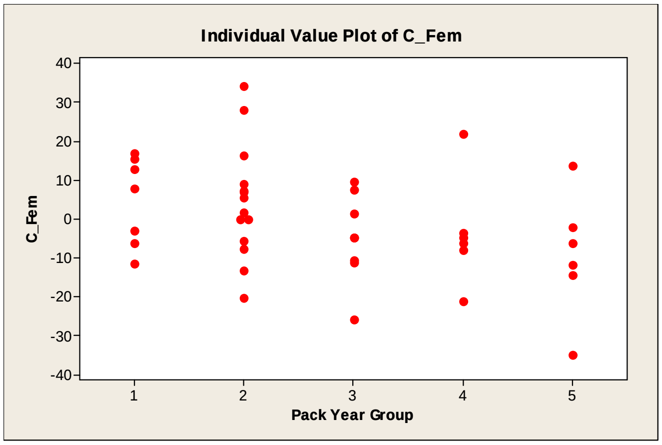
It appears that the value of C is generally decreasing as the difference in pack-years gets larger. This suggests that the lumbar spine bone mineral density is smaller in the heavier-smoking twin, which suggests that tobacco use has a negative relationship with bone mineral density.

2.41-2.43 For femoral neck BMD, we find . . .

A	B	C
-0.04	0.7	-5.714285714
-0.1	0.69	-14.49275362
0.01	0.635	1.57480315
0.05	0.665	7.518796992
-0.16	0.62	-25.80645161
-0.06	0.53	-11.32075472
-0.05	0.805	-6.211180124
-0.07	0.525	-13.33333333
0.12	0.71	16.90140845
-0.03	0.885	-3.389830508
0.04	0.72	5.555555556
-0.09	0.805	-11.18012422
………	……………	…………………
0.04	0.44	9.090909091
-0.05	0.665	-7.518796992
-0.03	0.635	-4.724409449
0.14	0.64	21.875
0.12	0.73	16.43835616
-0.09	0.765	-11.76470588
	Mean	-0.466252903
	Median	-2.941176471
	Sd	14.16185979

Descriptive Statistics: C_Fem

Variable	Pack Year Group	Mean	StDev	Median
C_Fem	1	4.68	11.38	7.87
	2	4.51	14.83	3.68
	3	-4.78	11.44	-4.76
	4	-3.56	14.05	-5.36
	5	-9.24	16.00	-8.99



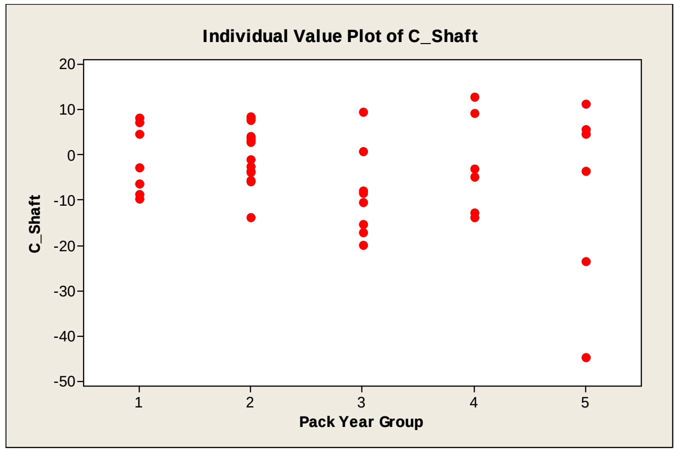
We get the same overall impression as before, that BMD decreases as tobacco use increases. The relationship may be a bit stronger using the femoral neck measurements, as we see a difference of approximately 14 units (4.68 – (-9.24)) in the mean value of C between Pack Year Group 1 (<10 py) and Pack Year Group 5 (>40 py). Using the lumbar spine data, this difference was approximately 11 units.

2.44-2.46 Using femoral shaft BMD, we find the following:

A	B	C
0.04	1.02	3.921568627
0.12	1.05	11.42857143
-0.19	0.955	-19.89528796
-0.09	1.075	-8.372093023
-0.18	1.05	-17.14285714
-0.07	1.095	-6.392694064
0.07	1.195	5.857740586
-0.01	1.045	-0.956937799
0.08	1.11	7.207207207
....	....	....
-0.1	1.17	-8.547008547
-0.08	1.01	-7.920792079
-0.03	0.875	-3.428571429
-0.04	0.68	-5.882352941
0.1	1.16	8.620689655
-0.2	1.32	-15.15151515
-0.03	1.045	-2.870813397
-0.04	1.04	-3.846153846
0.06	1.28	4.6875
	Mean	-3.241805211
	Median	-2.870813397
	Sd	11.29830441

Descriptive Statistics: C_Shaft

```
          Pack
          Year
Variable  Group    Mean  StDev  Median
C_Shaft   1       -0.98   7.67   -2.74
          2        0.25   6.49    1.03
          3       -8.55   9.77   -9.40
          4       -1.92  11.03   -3.80
          5       -8.26  21.61    0.63
```



When using the femoral shaft BMD data, the relationship between BMD and tobacco is much less clear. The lowest mean (and median) C value occurs in group 3, and it is hard to tell if any relationship exists between pack-year group and C.

2.47 We first read the data set LVM and show its first observations

```
> require(xlsx)
>lvm<-na.omit(read.xlsx("C:/Data_sets/lvm.xlsx", 1, header=TRUE))
> head(lvm)
  ID lvmht27 bpcat gender   age   BMI
1  1  31.281     1      1 17.63 21.45
2  2  36.780     1      2 16.11 19.78
3  6  20.660     1      2 17.03 20.58
4 10  44.222     1      2 11.50 25.34
5 16  23.302     1      1 11.90 17.30
6 20  27.735     1      2 10.47 19.16
```

We use the R function *tapply* to calculate the mean of LVMI by blood pressure group

```
> tapply(lvm$lvmht27, lvm$bpcat, mean)
       1        2        3
29.34266 33.79100 34.11569
```

2.48 We use also the R function *tapply* to calculate the geometric mean of LVMI by blood pressure group

```
> exp(tapply(log(lvm$lvmht27), lvm$bpcat, mean))
       1        2        3
28.60586 33.34814 32.88941
```

2.49
```
> boxplot(lvm$lvmht27~lvm$bpcat, main="Box plot of LVMI by blood
pressure group")
```

2.50 Since the box plots by blood pressure group are skewed, the geometric mean provides a more appropriate measure of location for this type of data.