Chapter 1

Introduction to Statistics and Data Analysis

1.1 (a) 15.

(b) $\bar{x} = \frac{1}{15}(3.4 + 2.5 + 4.8 + \cdots + 4.8) = 3.787$.

(c) Sample median is the 8th value, after the data is sorted from smallest to largest: 3.6.

(d) A dot plot is shown below.

(e) After trimming total 40% of the data (20% highest and 20% lowest), the data becomes:

2.9 3.0 3.3 3.4 3.6
3.7 4.0 4.4 4.8

So. the trimmed mean is

$$x_{tr20} = \frac{1}{9}(2.9+3.0+\cdots+4.8)=3.678.$$

(f) They are about the same.

1.2 (a) Mean=20.7675 and Median=20.610.

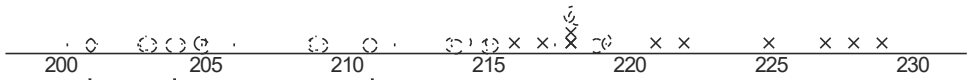
(b) $x_{tr10} = 20.743$.

(c) A dot plot is shown below.

(d) No. They are all close to each other.

1

1.3 (a) A dot plot is shown below.



In the figure, "×" represents the "No aging" group and "◦" represents the "Aging" group.

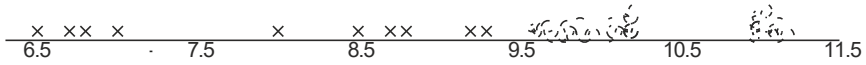
(b) Yes; tensile strength is greatly reduced due to the aging process.

(c) $\text{Mean}_{\text{Aging}} = 209.90$, and $\text{Mean}_{\text{Noaging}} = 222.10$.

(d) $\text{Median}_{\text{Aging}} = 210.00$, and $\text{Median}_{\text{Noaging}} = 221.50$. The means and medians for each group are similar to each other.

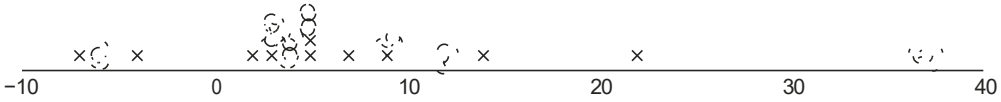
1.4 (a) $X_A = 7.950$ and $X_A = 8.250$;
$X_B = 10.260$ and $X_B = 10.150$.

(b) A dot plot is shown below.



In the figure, "×" represents company *A* and "◦" represents company *B*. The steel rods made by company *B* show more flexibility.

1.5 (a) A dot plot is shown below.

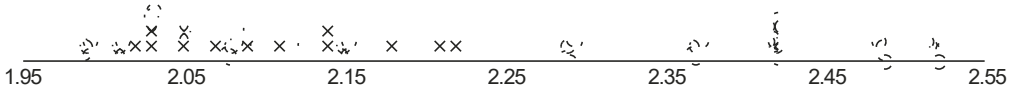


In the figure, "×" represents the control group and "◦" represents the treatment group.

(b) $X_{\text{Control}} = 5.60$, $X_{\text{Control}} = 5.00$, and $X_{\text{tr(10);Control}} = 5.13$;
$X_{\text{Treatment}} = 7.60$, $X_{\text{Treatment}} = 4.50$, and $X_{\text{tr(10);Treatment}} = 5.63$.

(c) The difference of the means is 2.0 and the differences of the medians and the trimmed means are 0.5, which are much smaller. The possible cause of this might be due to the extreme values (outliers) in the samples, especially the value of 37.

1.6 (a) A dot plot is shown below.



In the figure, "×" represents the 20◦C group and "◦" represents the 45◦C group.

(b) $X_{20^\circ\text{C}} = 2.1075$, and $X_{45^\circ\text{C}} = 2.2350$.

(c) Based on the plot, it seems that high temperature yields more high values of tensile strength, along with a few low values of tensile strength. Overall, the temperature does have an influence on the tensile strength.

(d) It also seems that the variation of the tensile strength gets larger when the cure temperature is increased.

1.7 $s^2 = \frac{1}{15-1}[(3.4 - 3.787)^2 + (2.5 - 3.787)^2 + (4.8 - 3.787)^2 + \cdots + (4.8 - 3.787)^2] = 0.94284$;
$s = \sqrt{s^2} = \sqrt{0.9428} = 0.971$.

1.8 $s^2 = \frac{1}{20-1}[(18.71 - 20.7675)^2 + (21.41 - 20.7675)^2 + \cdots + (21.12 - 20.7675)^2] = 2.5329;$

$s = \sqrt{2.5345} = 1.5915.$

1.9 (a) $s^2_{\text{No Aging}} = \frac{1}{\sqrt{-1}}[(227 - 222.10)^2 + (222 - 222.10)^2 + \cdots + (221 - 222.10)^2] = 23.66;$
$s_{\text{NoAging}} = \sqrt{23.62} = 4.86.$
$s^2_{\text{Aging}} = \frac{1}{\sqrt{-1}}[(219 - 209.90)^2 + (214 - 209.90)^2 + \cdots + (205 - 209.90)^2] = 42.10;$
$s_{\text{Aging}} = \sqrt{42.12} = 6.49.$

(b) Based on the numbers in (a), the variation in "Aging" is smaller that the variation in "No Aging" although the difference is not so apparent in the plot.

1.10 For company A: $s^2_A = 1.2078$ and $s_A = \sqrt{1.2078}$
For company B: $s^2_B = 0.3249$ and $s_B = \sqrt{0.3249} = 0.570.$

1.11 For the control group: $s^2_{\text{Control}} = 69.38$ and $s_{\text{Control}} = 8.33.$
For the treatment group: $s^2_{\text{Treatment}} = 128.04$ and $s_{\text{Treatment}} = 11.32.$

1.12 For the cure temperature at 20°C: $s^2_{20°C} = 0.005$ and $s_{20°C} = 0.071.$
For the cure temperature at 45°C: $s^2_{45°C} = 0.0413$ and $s_{45°C} = 0.2032$
The variation of the tensile strength is influenced by the increase of cure temperature.

1.13 (a) Mean $= \bar{X} = 124.3$ and median $= \tilde{X} = 120;$

(b) 175 is an extreme observation.

1.14 (a) Mean $= \bar{X} = 570.5$ and median $= \tilde{X} = 571;$

(b) Variance $= s^2 = 10$; standard deviation $= s = 3.162$; range $= 10;$

(c) Variation of the diameters seems too big so the quality is questionable.

1.15 Yes. The value 0.03125 is actually a P-value and a small value of this quantity means that the outcome (i.e., *HHHHH*) is very unlikely to happen with a fair coin.

1.16 The term on the left side can be manipulated to

$$\sum_{i=1} x_i - n\bar{x} = \sum_{i=1} x_i - n\sum_{i=1} x_i = 0,$$

which is the term on the right side.

1.17 (a) $\bar{X}_{\text{smokers}} = 43.70$ and $\bar{X}_{\text{nonsmokers}} = 30.32;$

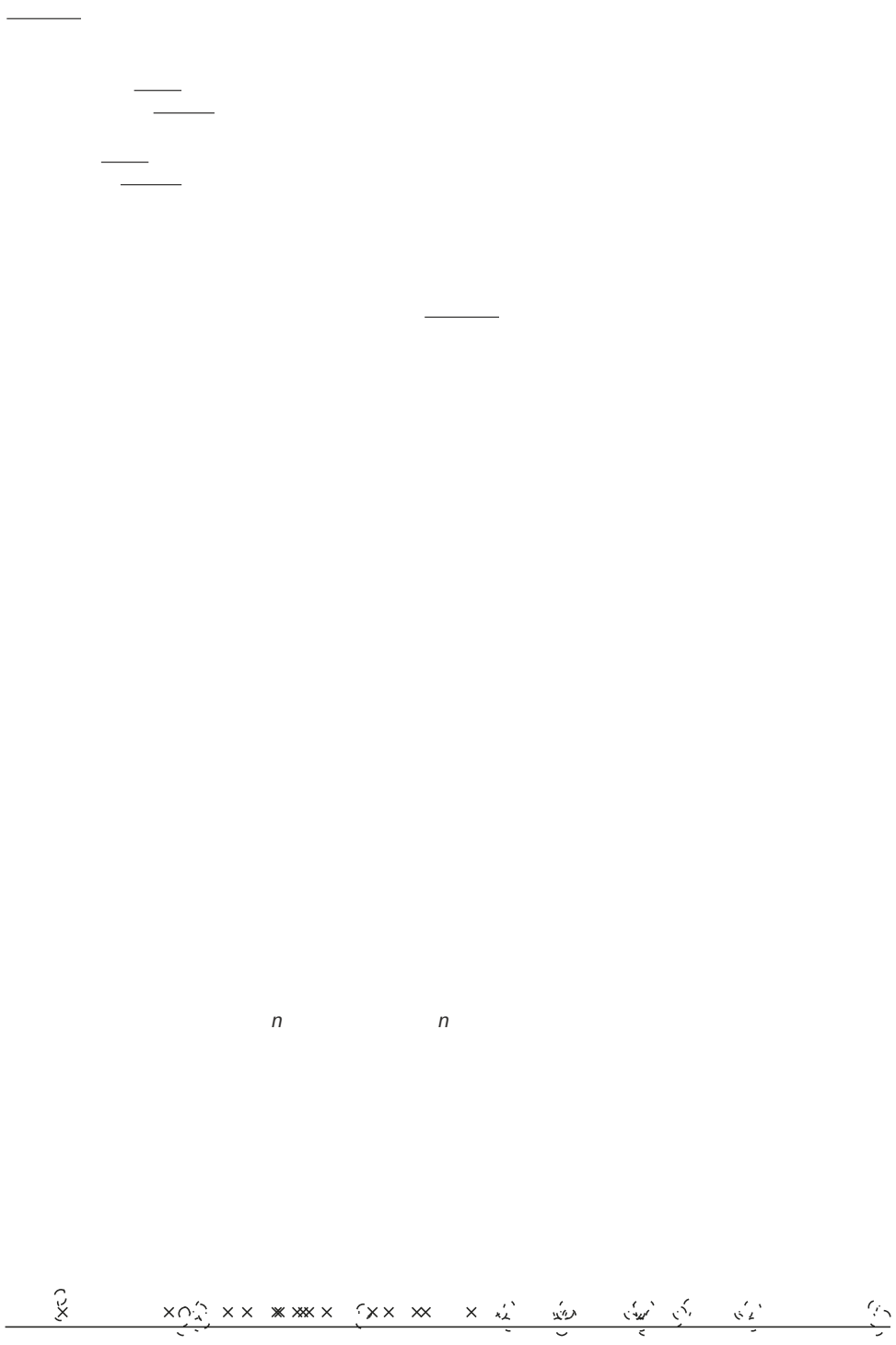
(b) $s_{\text{smokers}} = 16.93$ and $s_{\text{nonsmokers}} = 7.13;$

(c) A dot plot is shown below.

In the figure, "×" represents the nonsmoker group and "◦" represents the smoker group.

(d) Smokers appear to take longer time to fall asleep and the time to fall asleep for smoker group is more variable.

1.18 (a) A stem-and-leaf plot is shown below.



n *n*

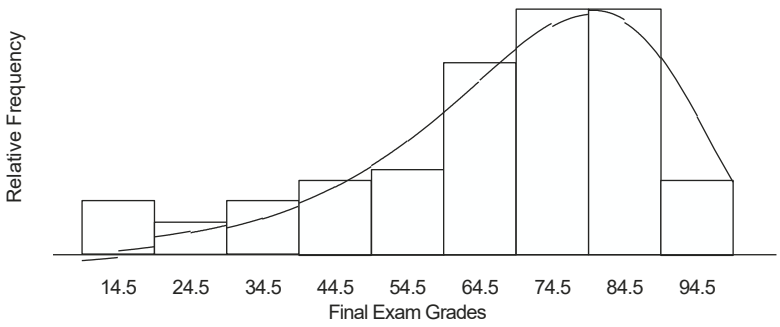
Stem	Leaf	Frequency
1	057	3
2	35	2
3	246	3
4	1138	4
5	22457	5
6	00123445779	11
7	01244456678899	14
8	00011223445589	14
9	0258	4

(b) The following is the relative frequency distribution table.

Relative Frequency Distribution of Grades

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
10 − 19	14.5	3	0.05
20 − 29	24.5	2	0.03
30 − 39	34.5	3	0.05
40 − 49	44.5	4	0.07
50 − 59	54.5	5	0.08
60 − 69	64.5	11	0.18
70 − 79	74.5	14	0.23
80 − 89	84.5	14	0.23
90 − 99	94.5	4	0.07

(c) A histogram plot is given below.



The distribution skews to the left.

(d) $\bar{X} = 65.48$, $\tilde{X} = 71.50$ and $s = 21.13$.

1.19 (a) A stem-and-leaf plot is shown below.

Stem	Leaf	Frequency
0	22233457	8
1	023558	6
2	035	3
3	03	2
4	057	3
5	0569	4
6	0005	4

(b) The following is the relative frequency distribution table.

Relative Frequency Distribution of Years

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
0.0 − 0.9	0.45	8	0.267
1.0 − 1.9	1.45	6	0.200
2.0 − 2.9	2.45	3	0.100
3.0 − 3.9	3.45	2	0.067
4.0 − 4.9	4.45	3	0.100
5.0 − 5.9	5.45	4	0.133
6.0 − 6.9	6.45	4	0.133

(c) $\bar{X}$ = 2.797, s = 2.227 and Sample range is 6.5 − 0.2 = 6.3.

1.20 (a) A stem-and-leaf plot is shown next.

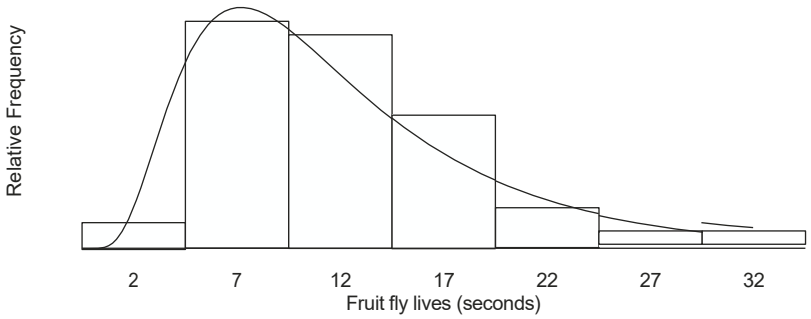
Stem	Leaf	Frequency
0*	34	2
0	56667777777889999	17
1*	0000001223333344	16
1	5566788899	10
2*	034	3
2	7	1
3*	2	1

(b) The relative frequency distribution table is shown next.

Relative Frequency Distribution of Fruit Fly Lives

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
0−4	2	2	0.04
5−9	7	17	0.34
10 − 14	12	16	0.32
15 − 19	17	10	0.20
20 − 24	22	3	0.06
25 − 29	27	1	0.02
30 − 34	32	1	0.02

(c) A histogram plot is shown next.



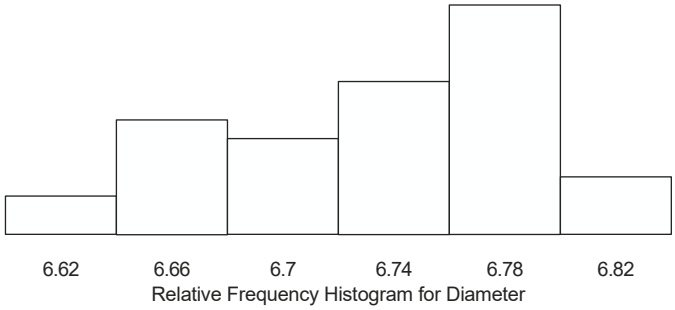
(d) $\bar{X}$ = 10.50.

1.21 (a) $\bar{X}$ = 74.02 and X = 78;

(b) s = 39.26.

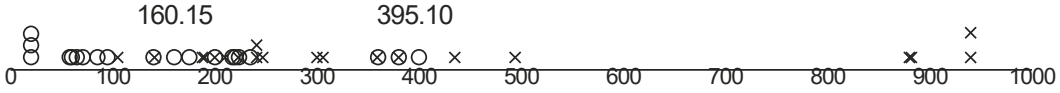
1.22 (a) $\bar{X}$ = 6.7261 and X = 0.0536.

(b) A histogram plot is shown next.



(c) The data appear to be skewed to the left.

1.23 (a) A dot plot is shown next.

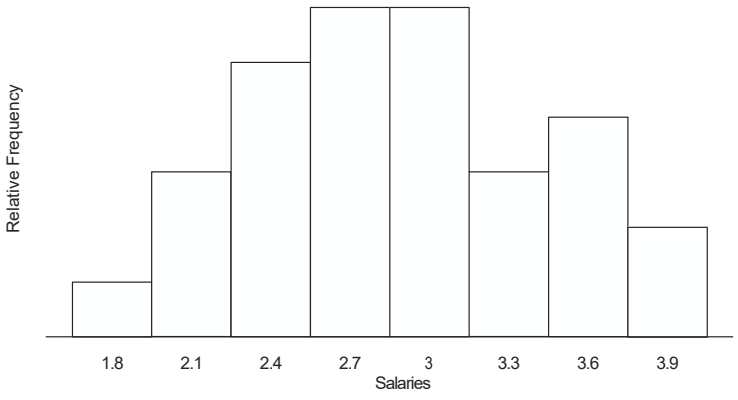


(b) X_{1980} = 395.1 and X_{1990} = 160.2.

(c) The sample mean for 1980 is over twice as large as that of 1990. The variability for 1990 decreased also as seen by looking at the picture in (a). The gap represents an increase of over 400 ppm. It appears from the data that hydrocarbon emissions decreased considerably between 1980 and 1990 and that the extreme large emission (over 500 ppm) were no longer in evidence.

1.24 (a) $\bar{X}$ = 2.8973 and s = 0.5415.

(b) A histogram plot is shown next.



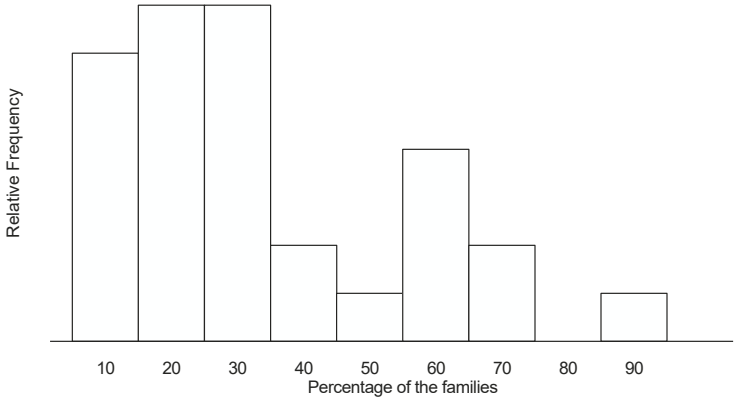
(c) Use the double-stem-and-leaf plot, we have the following.

Stem	Leaf	Frequency
1	(84)	1
2*	(05)(10)(14)(37)(44)(45)	6
2	(52)(52)(67)(68)(71)(75)(77)(83)(89)(91)(99)	11
3*	(10)(13)(14)(22)(36)(37)	6
3	(51)(54)(57)(71)(79)(85)	6

1.25 (a) $\bar{X} = 33.31$;

(b) $\bar{X} = 26.35$;

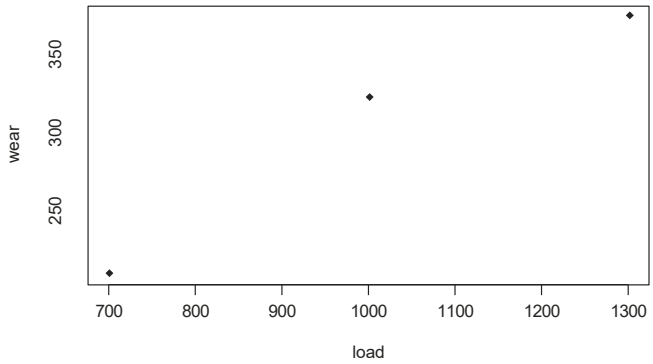
(c) A histogram plot is shown next.



(d) $\bar{X}_{tr(10)} = 30.97$. This trimmed mean is in the middle of the mean and median using the full amount of data. Due to the skewness of the data to the right (see plot in (c)), it is common to use trimmed data to have a more robust result.

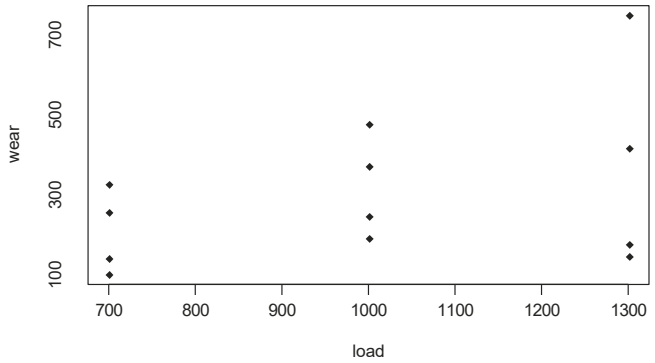
1.26 If a model using the function of percent of families to predict staff salaries, it is likely that the model would be wrong due to several extreme values of the data. Actually if a scatter plot of these two data sets is made, it is easy to see that some outlier would influence the trend.

1.27 (a) The averages of the wear are plotted here.



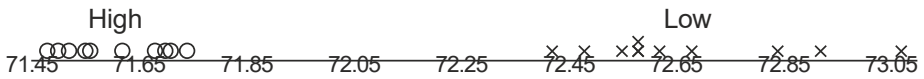
(b) When the load value increases, the wear value also increases. It does show certain relationship.

(c) A plot of wears is shown next.



(d) The relationship between load and wear in (c) is not as strong as the case in (a), especially for the load at 1300. One reason is that there is an extreme value (750) which influence the mean value at the load 1300.

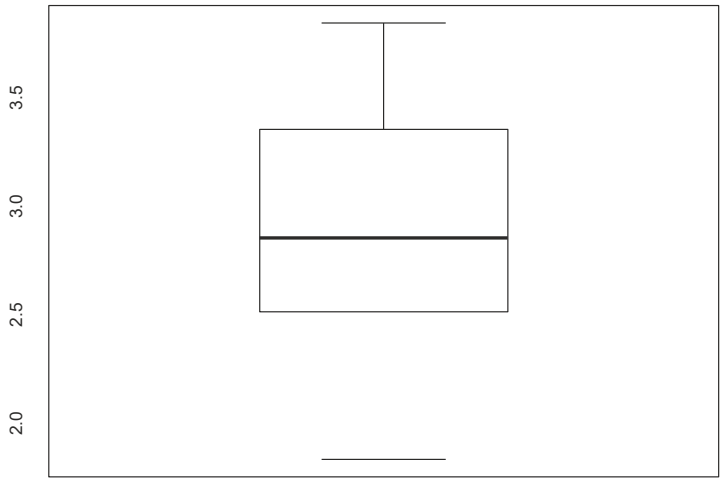
1.28 (a) A dot plot is shown next.



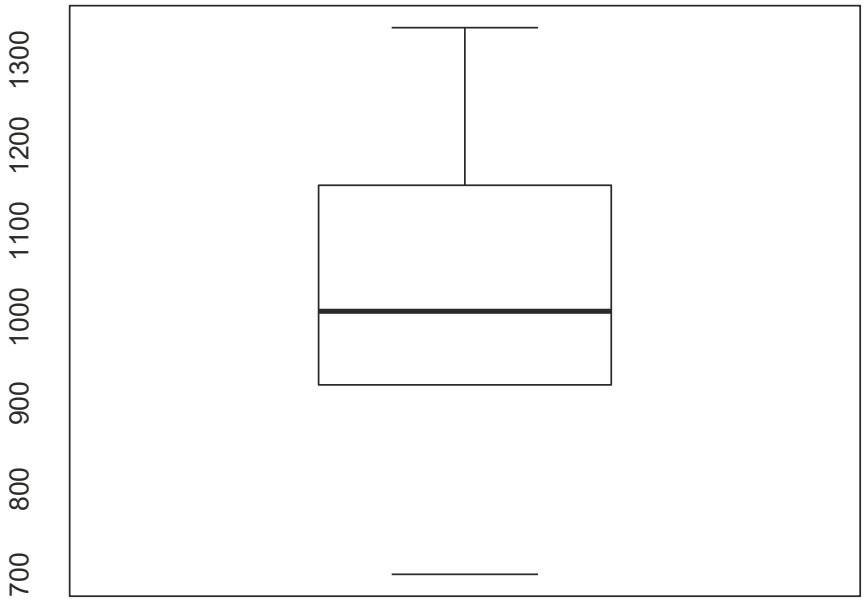
In the figure, "×" represents the low-injection-velocity group and "◦" represents the high-injection-velocity group.

(b) It appears that shrinkage values for the low-injection-velocity group is higher than those for the high-injection-velocity group. Also, the variation of the shrinkage is a little larger for the low injection velocity than that for the high injection velocity.

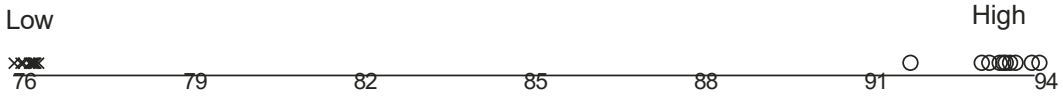
1.29 A box plot is shown next.



1.30 A box plot plot is shown next.



1.31 (a) A dot plot is shown next.

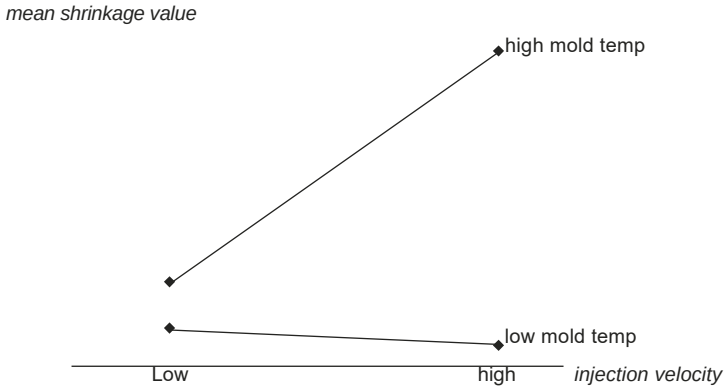


In the figure, "×" represents the low-injection-velocity group and "◦" represents the high-injection-velocity group.

(b) In this time, the shrinkage values are much higher for the high-injection-velocity group than those for the low-injection-velocity group. Also, the variation for the former group is much higher as well.

(c) Since the shrinkage effects change in different direction between low mode temperature and high mold temperature, the apparent interactions between the mold temperature and injection velocity are significant.

1.32 An interaction plot is shown next.



It is quite obvious to find the interaction between the two variables. Since in this experimental data, those two variables can be controlled each at two levels, the interaction can be inves-

tigated. However, if the data are from an observational studies, in which the variable values cannot be controlled, it would be difficult to study the interactions among these variables.