CHAPTER 1
Descriptive Statistics

Statistical software R is used for this book. All outputs and codes given are in R. R is a free statistical software, and it can be downloaded from the website: http://www.r-project.org

Exercises 1.2

1.2.1

The suggested solutions:

For qualitative data we can have color, sex, race, Zip code and so on. For quantitative data we can have age, temperature, time, height, weight and so on. For cross section data we can have school funding for each department in 2000. For time series data we can have the crude oil price from 1995 to 2008.

1.2.2

The suggested solutions:

For qualitative data we collect the frequency information of the data and we want to see the comparison by either bar chart or pie chart.

For quantitative data we collect the numerical information of the data and we want to see the comparison by histogram distribution.

For cross section data we collect different section data on the same time and we want to make comparison between them.

For time series data we collect same type of data on different time spot and we want to see if there is any trend or pattern of this data with time shifting.

1.2.3

The suggested questions can be:

1. What types of data the amounts are?
2. Do these Federal Agency receive the same amount of funding? If not, why?
3. Which Federal Agency should receive more funding? Why?

The suggested inferences we can make are:

1. These Federal Agency get different amount of money.
2. There are big differences between funding the Agencies receive.

1.2.4

The suggested questions can be

1. How does the funding changes for each agency through time?
2. Should we change the proportion between the Agencies or not?
3. Should we increase the total amount or not?

The suggested inferences we can make is

1. The total money tends to be the same.
2. The proportion between the Agencies tends to be the same.

Exercises 1.3

1.3.1

Simple Random Sample:

Say we have a population of 1,000 students, and we want a sample of 100 students. Using software or a random table, we randomly select 100 out of the 1,000 students. We want the selection probability for all the students to be equal. That is no student is more likely to be selected than any other student.

Systematic Sample:

Again, we have a population of 1,000 students, and we want a sample of 100 students. We need the sampling interval k = N/n = 10. Now, we need a random starting point between 1 and k. Let say, we randomly select 4. This gives us the sample: 4, 14, 24, ..., 74, 84, 94. This sample of numbers will correspond to ordered list of students.

Stratified Sample:

Suppose we decide to sample 100 college students from the population of 1000 (that is 10% of the population). We know these 1000 students come from three different major, Math, Computer Science and Social Science. We have Math 200, CS 400 and SS 400 students. Then we choose 10% of each of them Math 20, CS 40 and SS 40 by using simple random sample within each major.

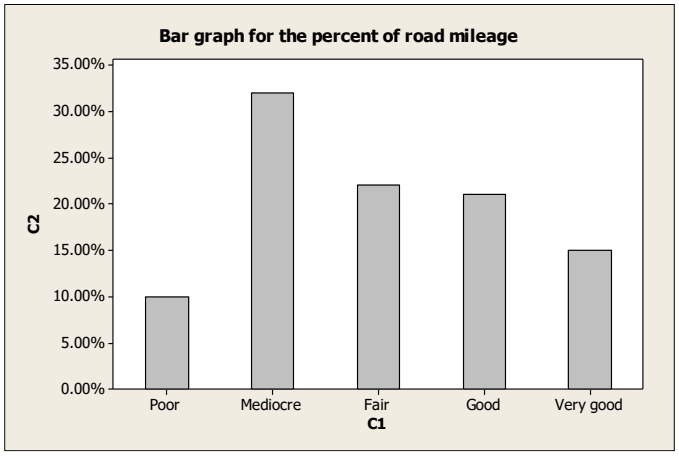
Cluster Sample:

Presume we have a population of 1,000 students clustered into 10 departments. For our sample of students, we will randomly select a subset from the 10 departments. Let say we randomly select 3 out 10 departments. Now, all the students on those 3 department become the sample from the population of students.

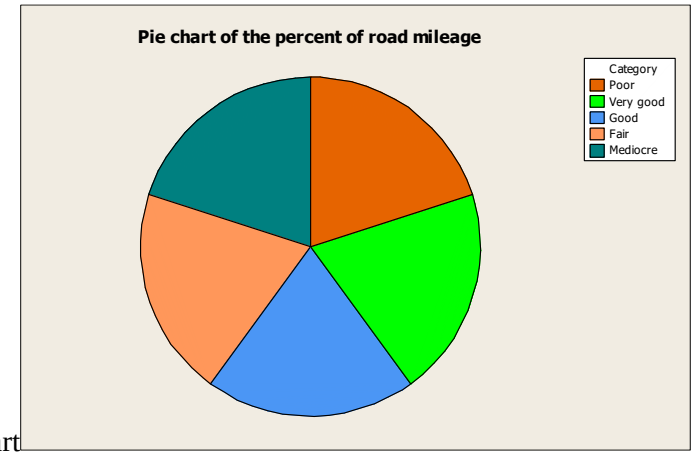
Exercises 1.4

1.4.1

(a) Bar graph

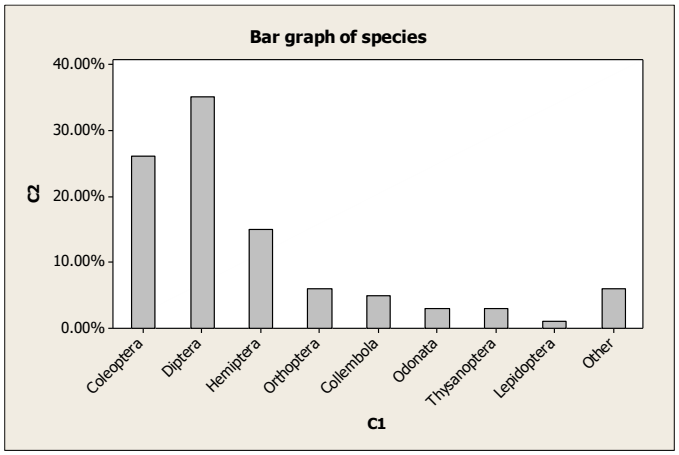


(b) Pie chart

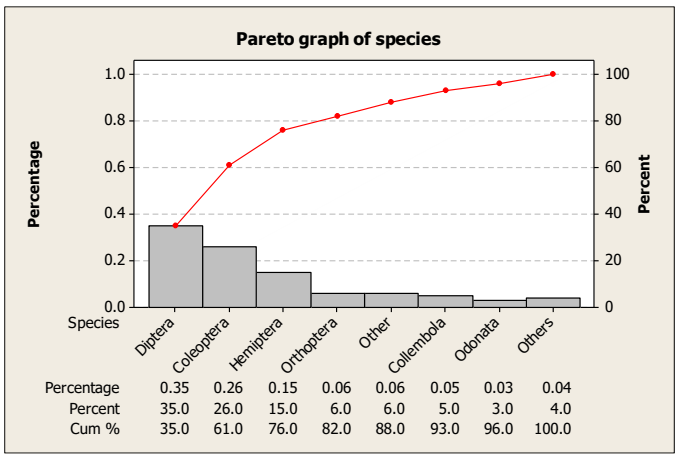


1.4.2

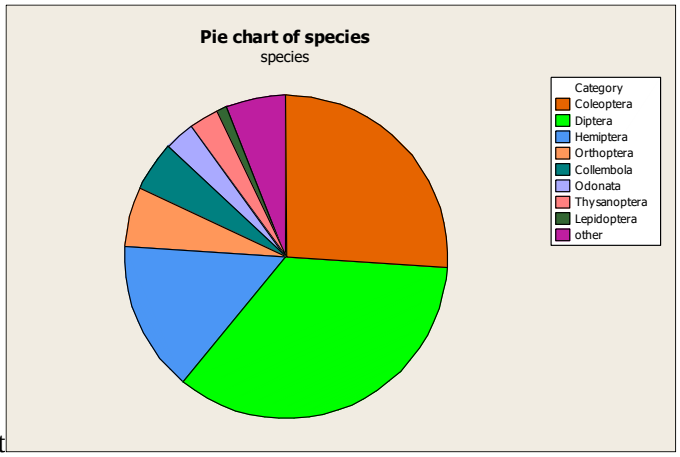
(a) Bar graph



(b) Pareto graph

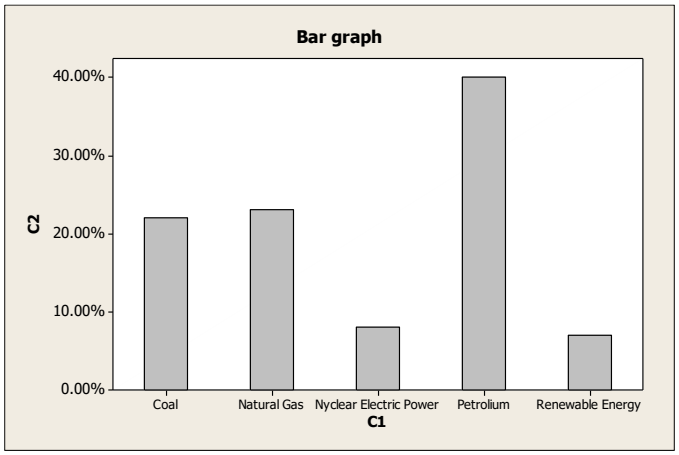


(c) Pie chart

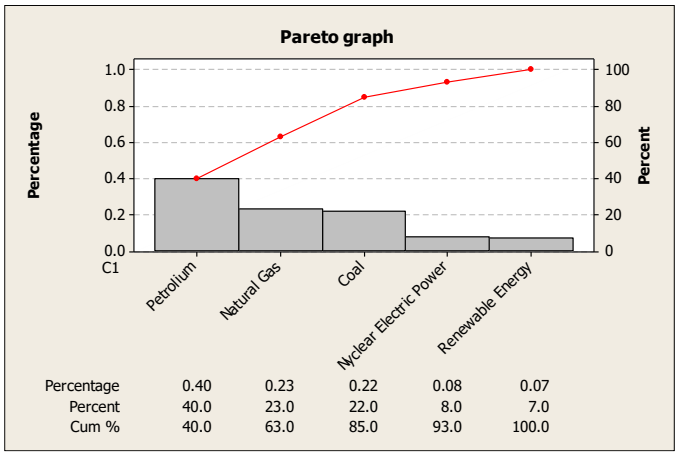


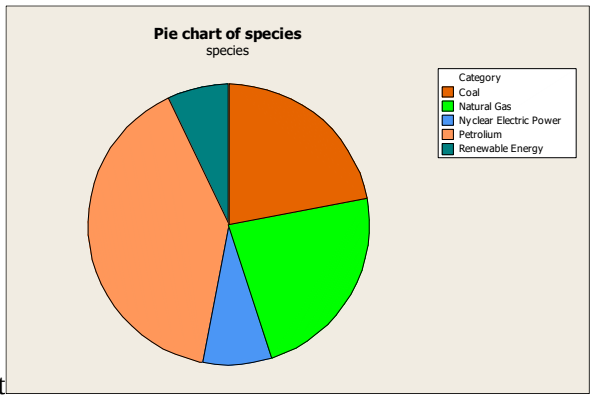
1.4.3

(a) Bar graph



(b) Pareto graph

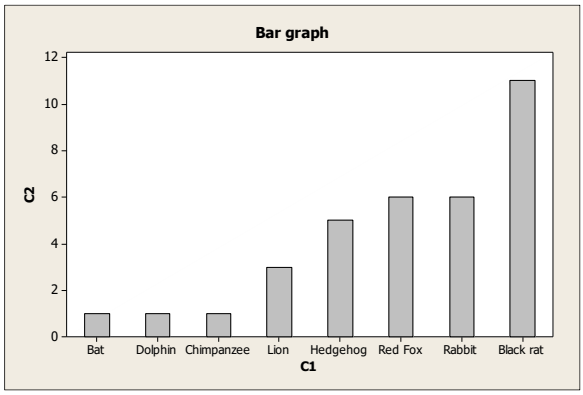




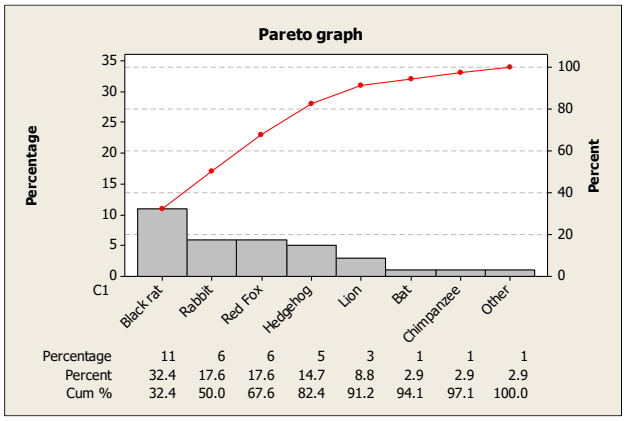
(c) Pie chart

1.4.4

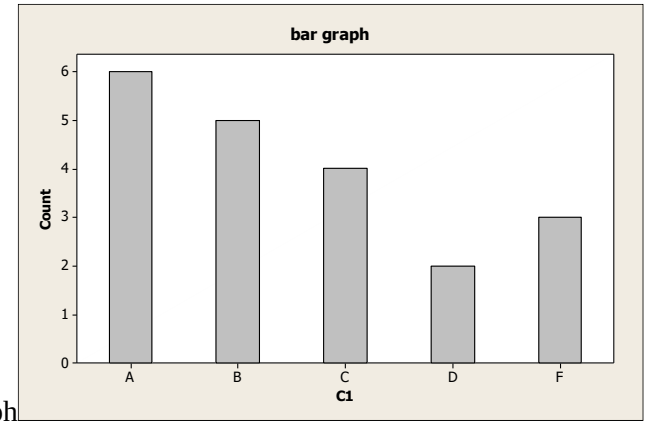
(a) Bar graph



(b) Pareto graph

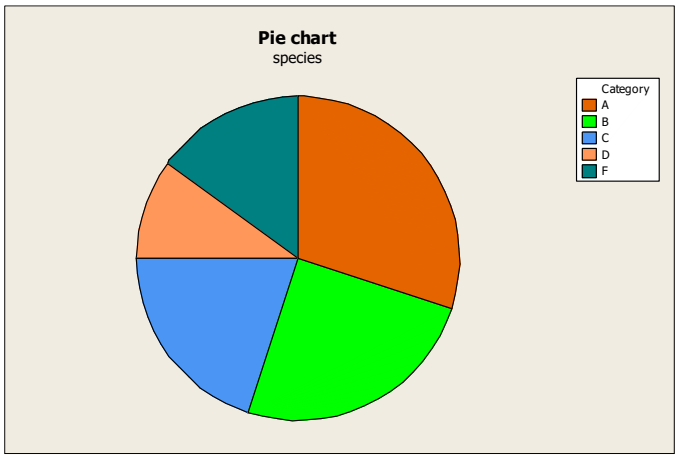


1.4.5



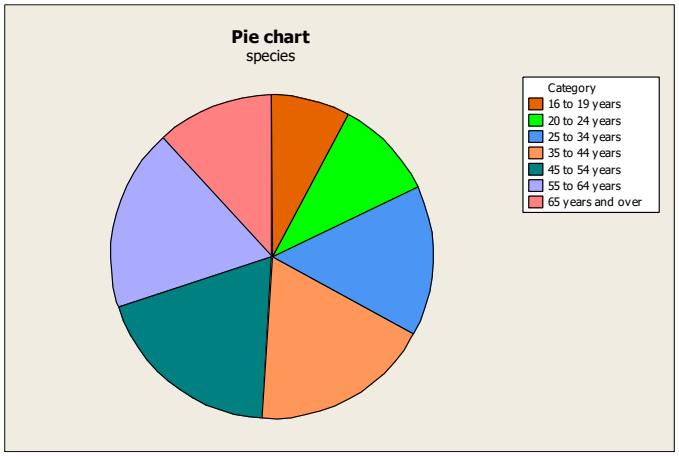
(a) Bar graph

(b) Pie chart

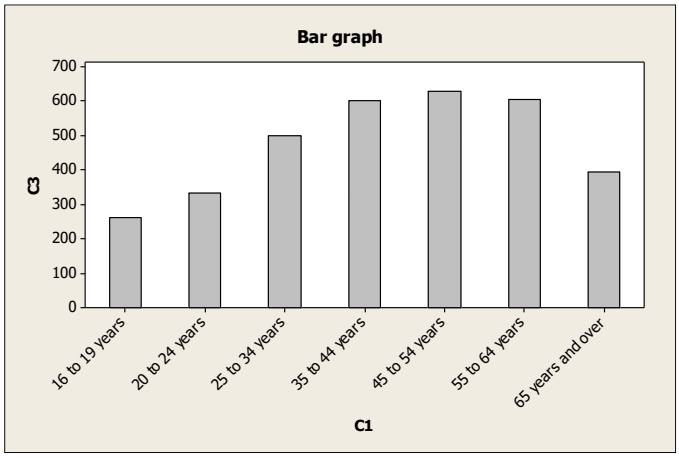


1.4.6

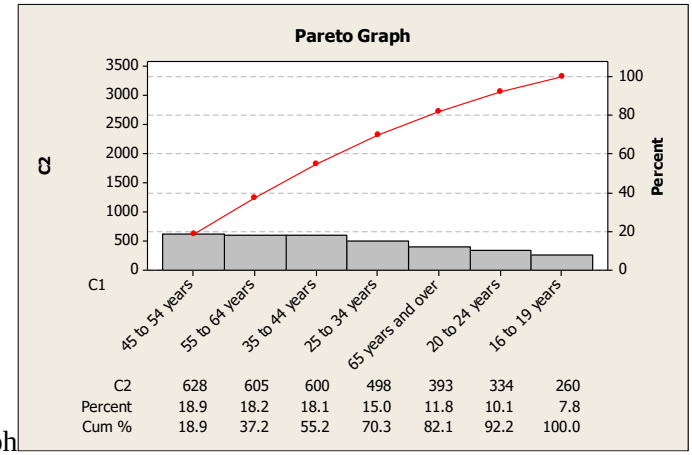
(a) Pie chart



(b) Bar graph

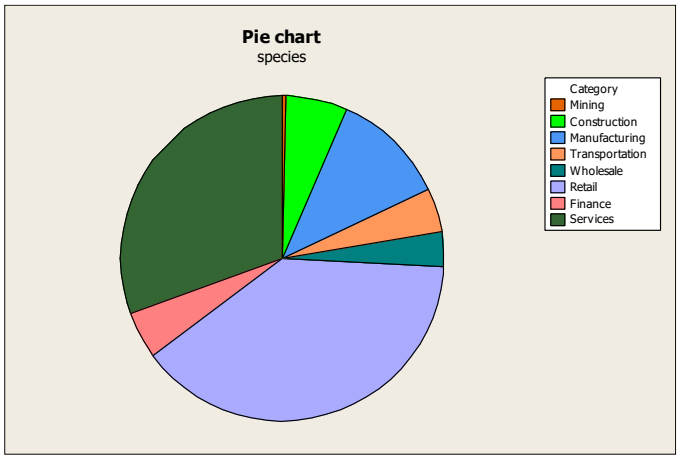


(c) Pareto graph

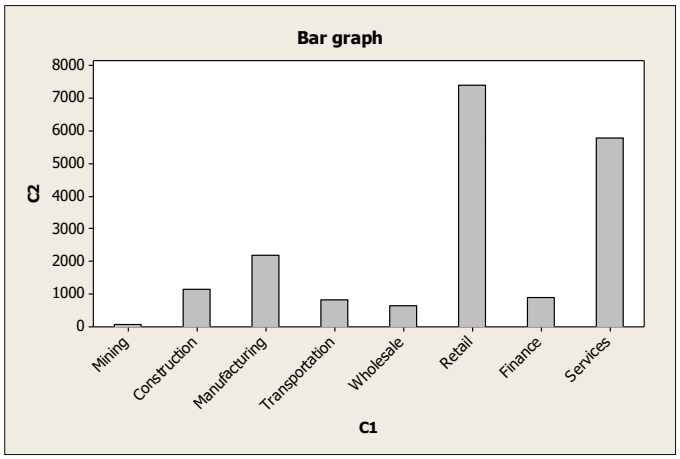


1.4.7

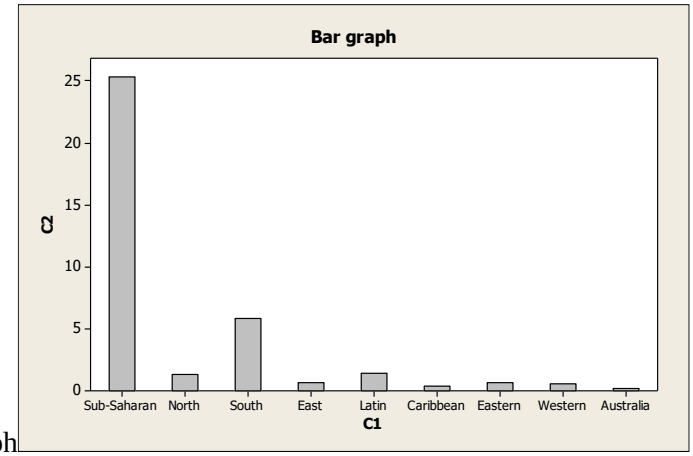
(a) Pie chart



(b) Bar graph

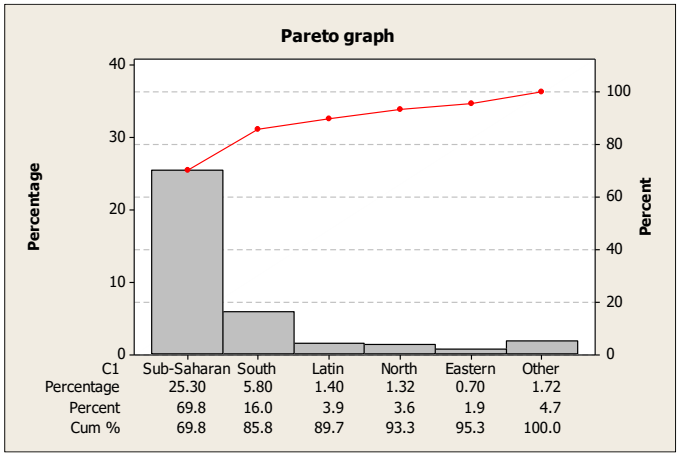


1.4.8



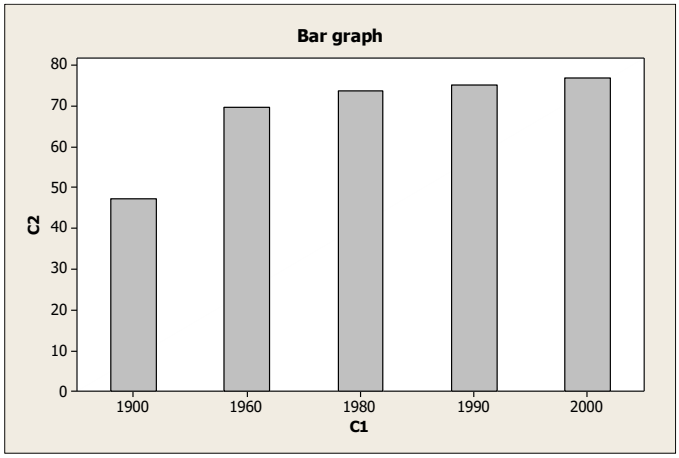
(a) Bar graph

(b) Pareto graph

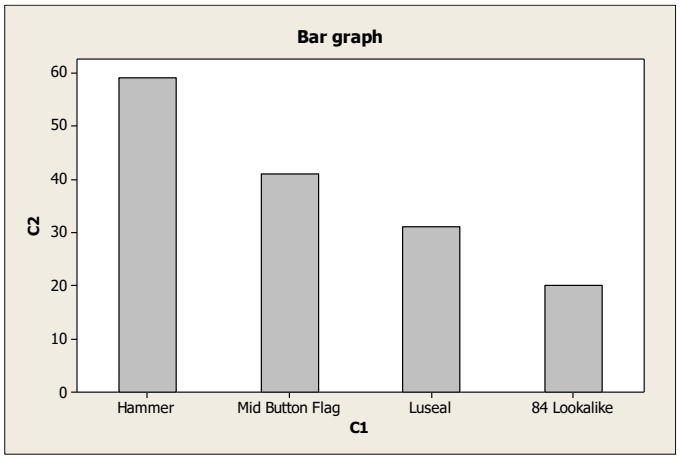


1.4.9

Bar graph

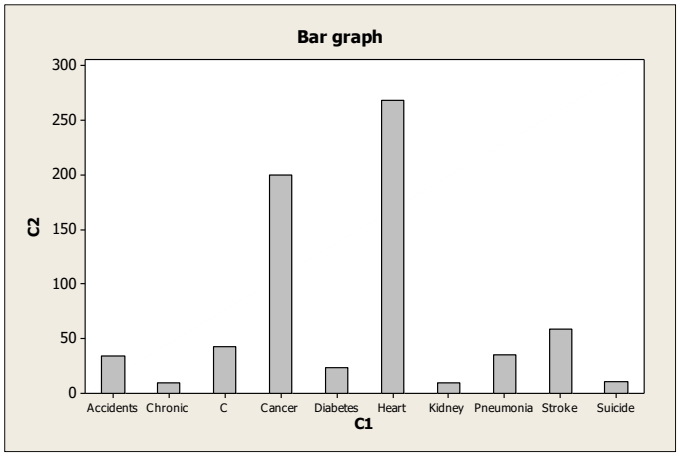


1.4.10

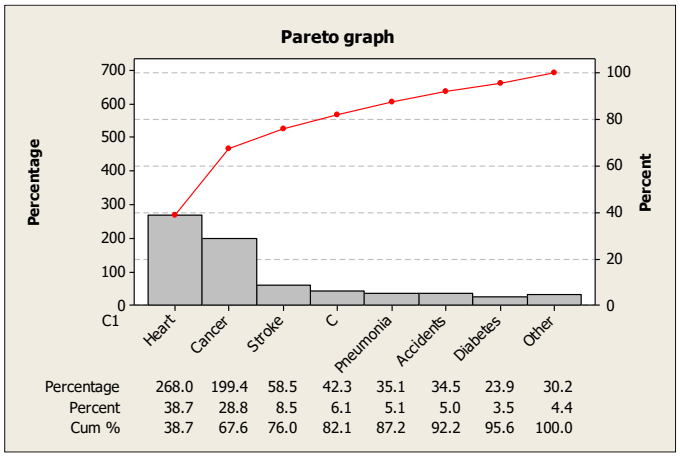


1.4.11

(a) Bar graph



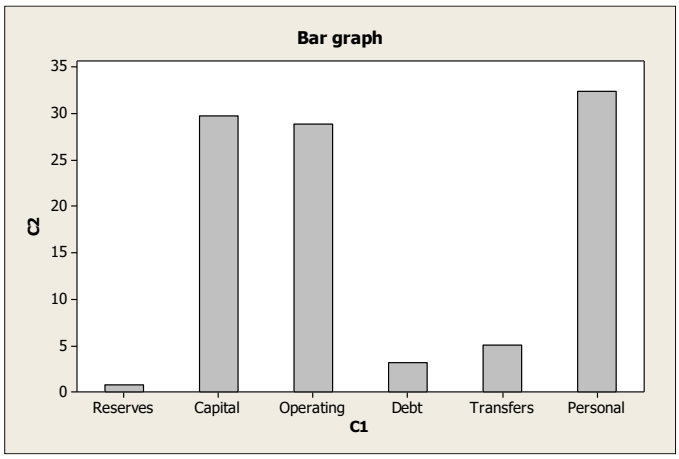
(b) Pareto graph



1.4.12

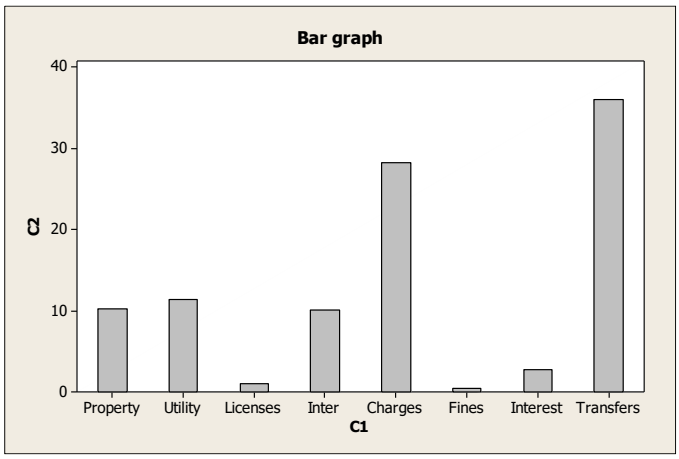
(a) Expenditure

Bar graph



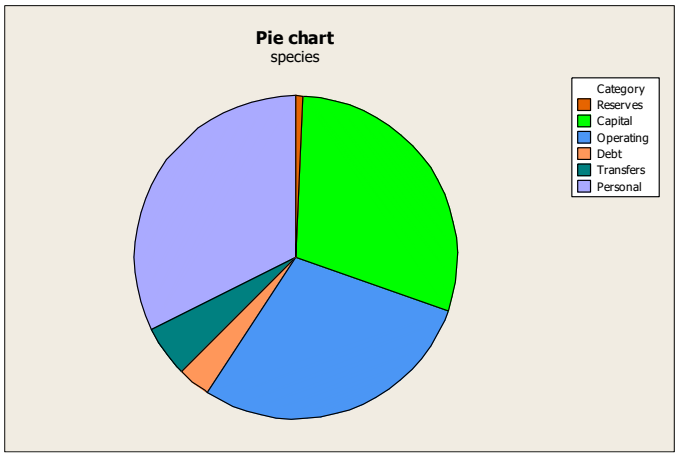
Revenues

Bar graph



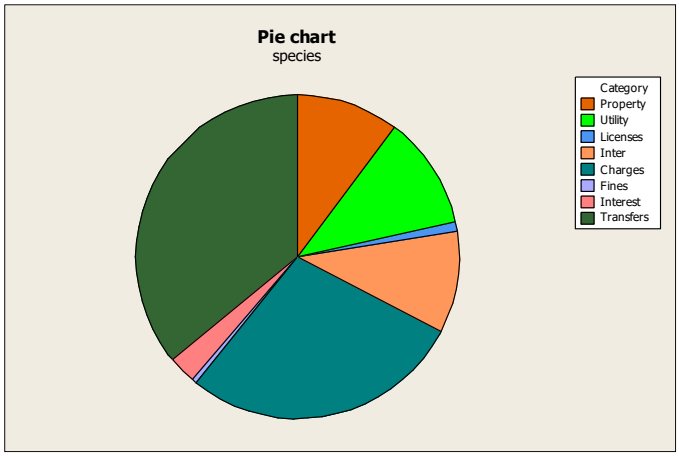
(b) Expenditure

Pie chart

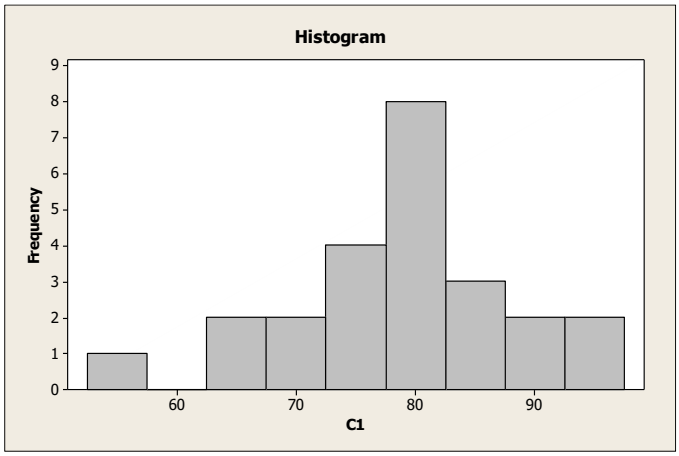


Revenues

Pie chart



1.4.13



1.4.14

(a) Stem and leaf

Stem-and-Leaf Display: C1

Stem-and-leaf of C1 N = 40

Leaf Unit = 1.0

2 0 00

12 0 2222223333

13 0 5

20 0 6666677

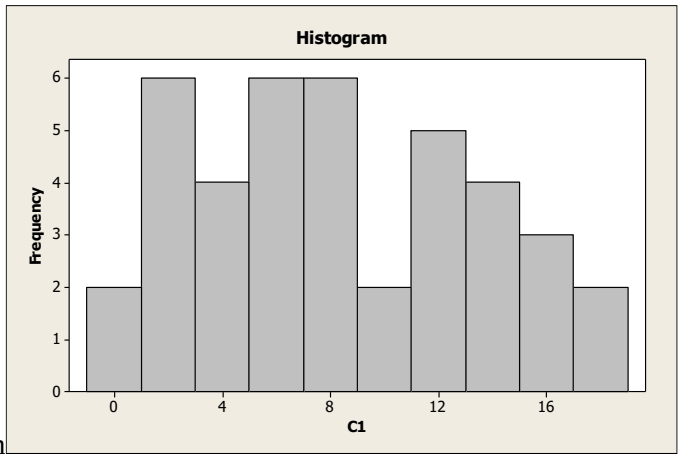
20 0 888899

14 1 111

11 1 223333

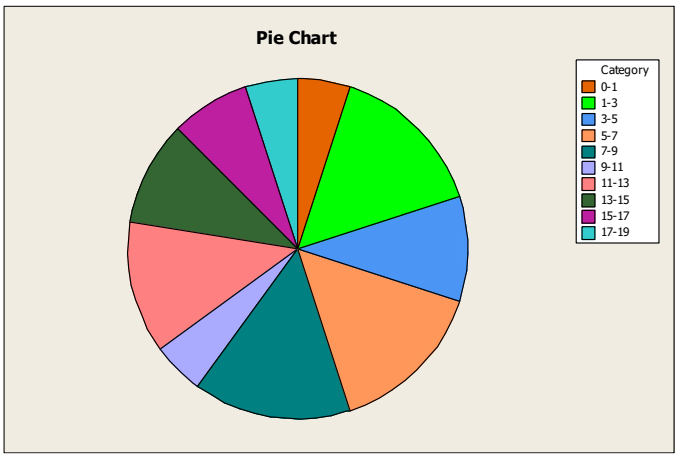
5 1 55

3 1 677



(b) Histogram

(c) Pie chart



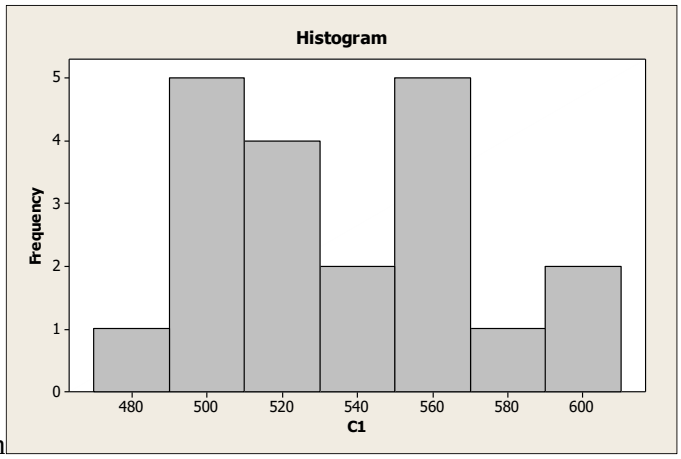
1.4.15

(a) Stem and leaf

Stem-and-leaf of SAT Mathematics scores N = 20

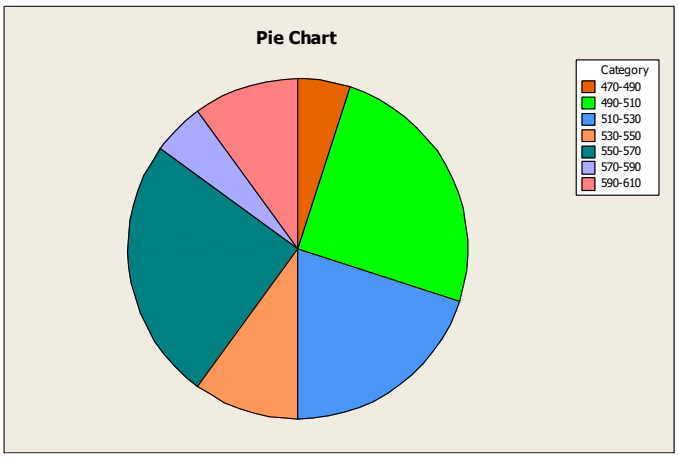
Leaf Unit = 10

1 4 7
3 4 99
8 5 00011
10 5 22
10 5 4455
6 5 6667
2 5 9
1 6 0



(b) Histogram

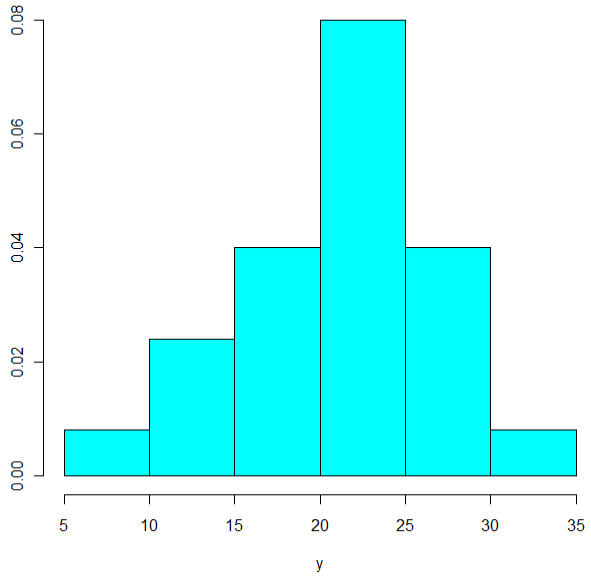
(c) Pie chart



1.4.16

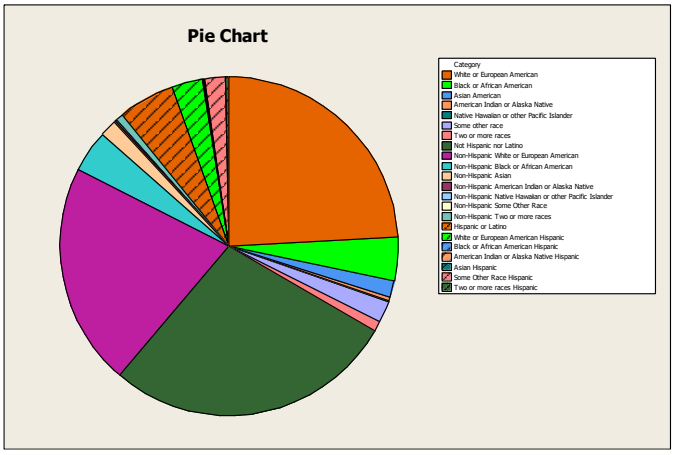
Frequency table

Interval	Frequency	Relative Freq	Percentage
5-9	1	.04	4
10-14	3	.12	12
15-19	5	.2	20
20-24	10	.4	40
25-29	5	.2	20
30-35	1	.04	4



Histogram

1.4.17



Exercises 1.5

1.5.1

$$\bar{x} = \frac{\sum f \cdot x}{n} = \frac{\quad}{\quad} = \quad$$

$$s^2 = \frac{\sum f \cdot x^2 - \frac{(\sum f \cdot x)^2}{n}}{n-1}$$

$$s^2 = \quad$$

$$s = \sqrt{\quad} = \quad$$

1.5.2

(a)

$$\bar{x} = \frac{7.625 + 7.5 + \cdots + 5.375 + 7.5}{10} = 7.03$$

$$s^2 = \frac{(7.625 - 7.03)^2 + (7.5 - 7.03)^2 + \cdots + (5.375 - 7.03)^2 + (7.5 - 7.03)^2}{10 - 1}$$

$$s^2 = .743$$

$$s = \sqrt{.743} = .8618$$

(b)

$$Q_1 = 6.625$$

$$Q_3 = \frac{7.5 + 7.625}{2} = 7.5625$$

$$M = 7.375$$

$$IQR = 7.5625 - 6.625 = .9375$$

$$LL = 6.625 - 1.5(.9375) = 5.21875$$

$$LL = 7.625 + 1.5(.9375) = 9.03125$$

There are no outliers.

1.5.3

Given information: mean=6 , median = 4 , mode = 3

We know that the value 3 can only be in the data twice. If not the median would be different than 4. This give us the following: 3, 3, x, y. Where x and y are the missing values. We introduce a system of equation to solve for x and y.

$$\frac{3 + 3 + x + y}{4} = 6 \qquad \frac{x + 3}{2} = 4$$

$$x + y = 24 - 6 \qquad x = 8 - 3$$

$$x + y = 18 \qquad x = 5$$

$$y = 18 - 5$$

$$y = 13, x = 5$$

Data: 3, 3, 5, 13

$$Var = \frac{1}{3}\left[(3-6)^2 + (3-6)^2 + (5-6)^2 + (13-6)^2\right]$$

$$= \frac{1}{3}(68)$$

$$= 22.667$$

$$Sd = \sqrt{Var}$$

$$= \sqrt{22.667}$$

$$= 4.76$$

1.5.4

$$\bar{x} = \frac{1183 + 1050 + \cdots + 1578 + 2621}{14} = 1245$$

$$s^2 = \frac{(1183-1245)^2 + (1050-1245)^2 + \cdots + (1578-1245)^2 + (2621-1245)^2}{14-1}$$

(a)

$$s^2 = 722653.1$$

$$s = \sqrt{722653.1} = 850.1$$

$$Range = 2621 - 537 = 2084$$

$Q_1 = 537$

$Q_3 = 1578$

$M = \dfrac{1117 + 1050}{2} = 1083.5$

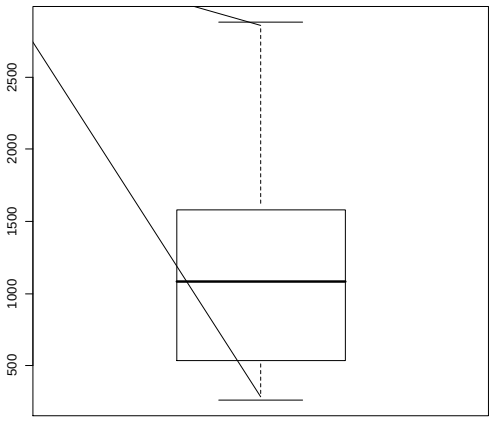
(b) $IQR = 1578 - 537 = 1041$

$LL = 537 - 1.5(1041) = -1024.5$

$LL = 1578 + 1.5(1041) = 3139.5$

There are no outliers.

(c)



1.5.5

$Q_1 = 80$

$Q_3 = 115$

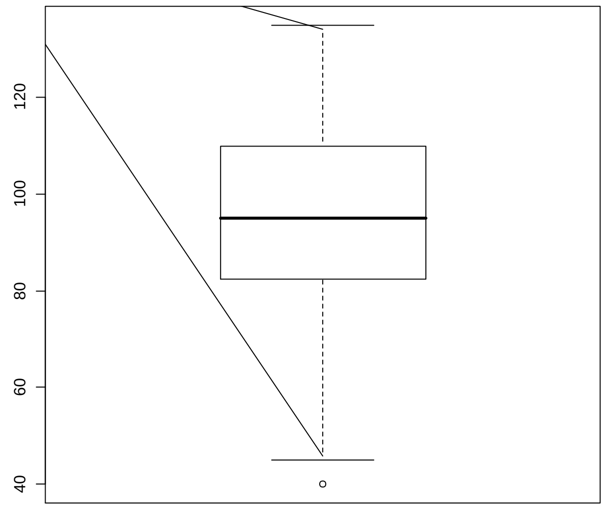
(a) $M = 95$

$IQR = 115 - 80 = 35$

$LL = 80 - 1.5(35) = 27.5$

$LL = 115 + 1.5(35) = 167.5$

(b)



(c) There are no outliers.

1.5.6

$$\bar{x} = \frac{5(2)+4(7)+15(12)+10(17)+6(22)}{50} = 13$$

$$s^2 = \frac{5(2-13)^2+4(7-13)^2+15(12-13)^2+10(17-13)^2+6(22-13)^2}{50-1} \approx [\text{illegible}]$$

$s = \sqrt{[\text{illegible}]}$

1.5.7

(a) [illegible]

(b) $\sum_{i=1}^{5} (x_i - \bar{x}) = [\text{illegible}]$

$= [\text{illegible}]$

$= [\text{illegible}]$

1.5.8

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2$$

(a) $$=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2$$

$$=\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}$$

(b)

$$\sum_{i=1}^{n}x_i^2=(10.24)^2+(6.24)^2+\cdots+(11.52)^2+(9.24)^2=9\text{X}.$$

$$\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}=1\text{X}9.(18)^2=9\text{X}.3$$

1.5.9

(a)
$$\bar{x}=\frac{\sum_{i=1}^{32}x_i}{32}=\frac{1056}{32}=33.05$$

$$s^2=\frac{1}{31}\sum_{i=1}^{32}(x_i-33.05)^2\approx\frac{548.32}{31}\approx17.04$$

$$range=53.0-53.1=48.9$$

$$Q_1=\frac{24.75+25.44}{2}=25.095$$

$$Q_3=\frac{42.19+43.25}{2}=42.72$$

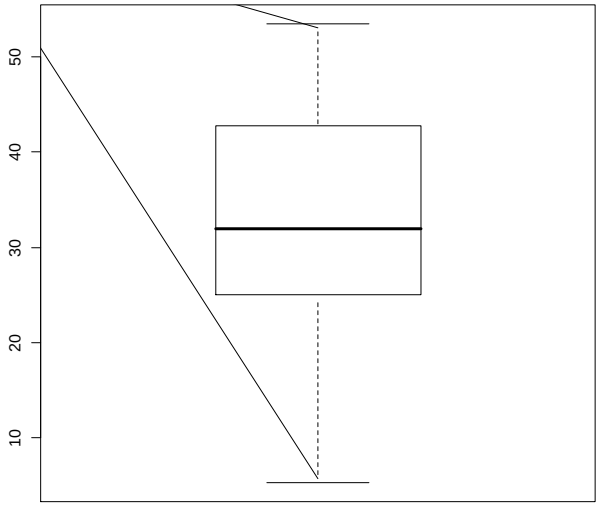
(b) $$M=\frac{32+32}{2}=32$$

$$IQR=42.72-25.095=17.625$$

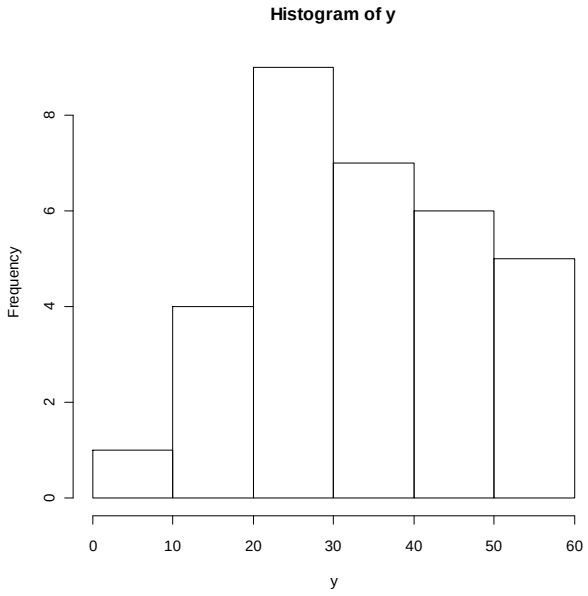
$$LL=25.095-1.5(17.625)=-1.3425$$

$$LL=42.72+1.5(17.625)=69.1575$$

There are no outliers.



(c)



(d)

(e)

$\bar{x} = 33.105$

$\bar{x} \pm s = [26.64]$ 21 data point (65.625%) fall within 1 SD, empirical rule = 68%

$\bar{x} \pm 2s = [49.52]$ 31 data point (96.875%) fall within 2 SD, empirical rule = 95%

$\bar{x} \pm 3s = [63.72]$ 32 data point (100%) fall within 3 SD, empirical rule = 99.7%

1.5.10

(a)
$$\bar{x} = \frac{\sum_{i=1}^{40} x_i}{40} = \frac{3336}{40} = 83.4$$

$$s^2 = \frac{1}{39}\sum_{i=1}^{40}(x_i - 83.4)^2 \approx \frac{944.36}{39} \approx 24.215$$

$$range = 172 - 5 = 167$$

$$Q_1 = \frac{3.7 + 3.6}{2} = 3.65$$

$$Q_3 = \frac{12.8 + 12.3}{2} = 12.55$$

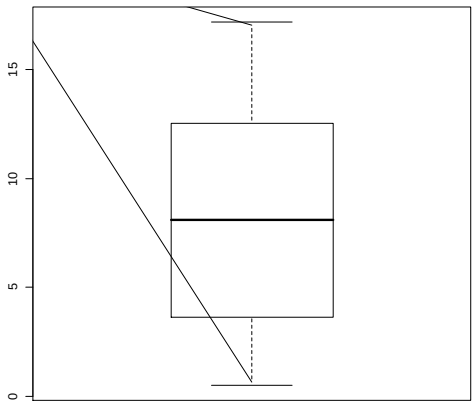
(b)
$$M = \frac{8.3 + 7.9}{2} = 8.1$$

$$IQR = 12.55 - 3.65 = 8.9$$

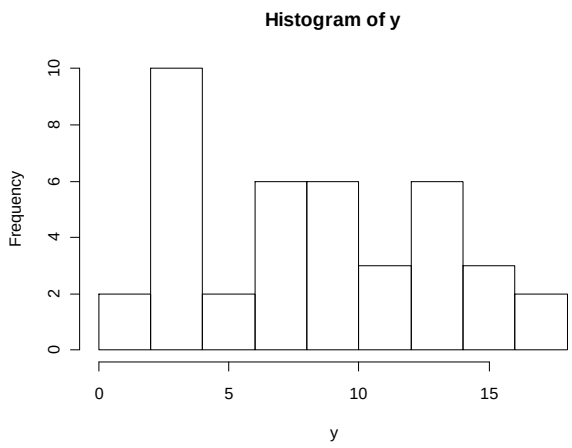
$$LL = 3.65 - 1.5(8.9) = -9.7$$

$$LL = 12.55 + 1.5(8.9) = 25.9$$

There are no outliers.



(c)



(d)

(e) $\bar{x} = 8.34$

$\bar{x} \pm s = [$ [illegible] $]$ 24 data point (60%) fall within 1 SD, empirical rule = 68%

$\bar{x} \pm 2s = [$ [illegible] $]$ 40 data point (100%) fall within 2 SD, empirical rule = 95%

$\bar{x} \pm 3s = [$ [illegible] $]$ 40 data point (100%) fall within 3 SD, empirical rule = 99.7%

1.5.11

(a) [illegible]

(b) [illegible]

1.5.12

$Q_1 = 39$

$Q_3 = 46$

$M = \dfrac{42 + 40}{2} = 41$

$IQR = 46 - 39 = 7$

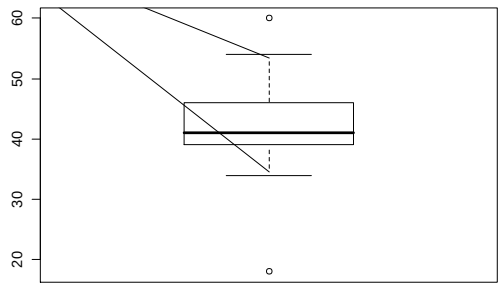
(a)

$\bar{x} = \dfrac{\sum_{i=1}^{10} x_i}{10} = \dfrac{418}{10} = 41.8$

$s^2 = \dfrac{1}{9}\sum_{i=1}^{10}\left(x_i - 8.34\right)^2 = \dfrac{1149.6}{9} \approx 127.733$

$s = \sqrt{127.733} \approx 11.302$

(b) [illegible]



(c)

$IQR = 7$

$IL = 39 - 1.5(7) = 28.5$

(d) $IL = 46 - 1.5(7) = 35.5$

There are two outliers: 18 and 60.

1.5.13

(a)

$$\bar{x} = \frac{\sum_{i=1}^{30} x_i}{30} = \frac{1123}{30} = 37.433$$

$$s^2 = \frac{1}{29}\sum_{i=1}^{30}(x_i - 37.433)^2 \approx 3.502$$

$$s = \sqrt{3.502} \approx 1.871$$

(b) Frequency table

Class	Interval	Frequency	M_i	$M_i \cdot f_i$
1	0-1.6	4	.8	3.2
2	1.7-3.3	10	2.5	25
3	3.4-5	9	4.2	37.8
4	5.1-6.7	5	5.9	29.5
5	6.8-8.4	2	7.6	15.2

(c) Grouped data:

$$\bar{x} = \frac{(0.8)(4) + (2.5)(10) + (4.2)(9) + (5.9)(5) + (7.6)(2)}{30} = 3.6$$

$$s^2 = \frac{1}{29}\left[(0.8-3.6)^2(4) + (2.5-3.6)^2(10) + (4.2-3.6)^2(9) + (5.9-3.6)^2(5) + (7.6-3.6)^2(2)\right] = 3.$$

$$s = \sqrt{3.319} = 1.822$$

The results from the grouped data are similar to the actual data.

1.5.14

(a)

$$\bar{x} = \frac{\sum_{i=1}^{30} x_i}{30} = \frac{1814}{30} \approx 60.467$$

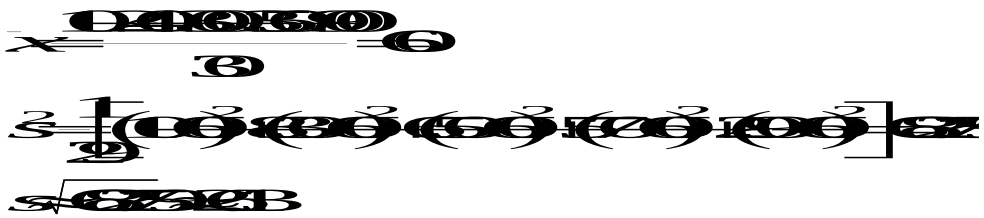
$$s^2 = \frac{1}{29}\sum_{i=1}^{30}(x_i - 60.467)^2 \approx 685.085$$

$$s \approx \sqrt{685.085} \approx 26.174$$

(b) Frequency table

Class	Interval	Frequency	M_i	$M_i \cdot f_i$
1	0-20	1	10	10
2	20-40	8	30	240
3	40-60	6	50	300
4	60-80	5	70	350
5	80-100	10	90	900

(c)
Grouped data:



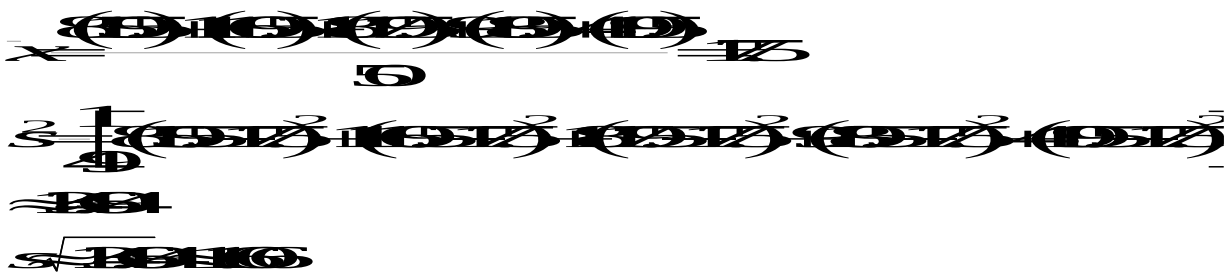
The results from the grouped data are similar to the actual data.

1.5.15

$L = 25$ $f_m = 1396$ $w = 4$

$F_b = 1788$ $n = 5146$

1.5.16
(a)



(b) $L = 175$, $f_m = 18$, $w = 9$

$F_b = 19$, $n = 50$

1.5.17

(b) $L = 40$, $f_m = 59$, $w = 19$, $F_b = 69$