Chapter 4: Summarizing Data Collected in the Sample

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Data

The **population** is the collection of all individuals about whom we wish to make generalizations.

Example: We wish to assess the prevalence of CVD among all adults aged 30 to 75 in the US.

The **sample** is a subset of individuals from the population.

Example: A researcher randomly selects 1000 adults aged 30 to 75 in the US to assess the prevalence of CVD.

Data

Dichotomous variables have only two possible responses. Yes/No

Example: Exposure to a risk factor such as smoking can be coded as yes or no. (Sometimes 1/0).

Ordinal variables have more than two possible ordered responses

Example: Symptom severity of minimal, moderate, and severe.

Data

Categorical variables sometimes called nominal variables are similar to ordinal variables except that the responses are unordered.

Example: Race/ethnicity.

Continuous variables take on an unlimited number of responses between defined minimum and maximum values.

Examples: Systolic blood pressure, diastolic blood pressure, total cholesterol level, CD4 count, platelet count, age, height, and weight.

Data

Statistics: Numerical summary measures computed on samples.

Example: The mean blood pressure among a random sample of 1000 adults aged 30 to 75 in the US.

Parameters: Summary measures computed on populations.

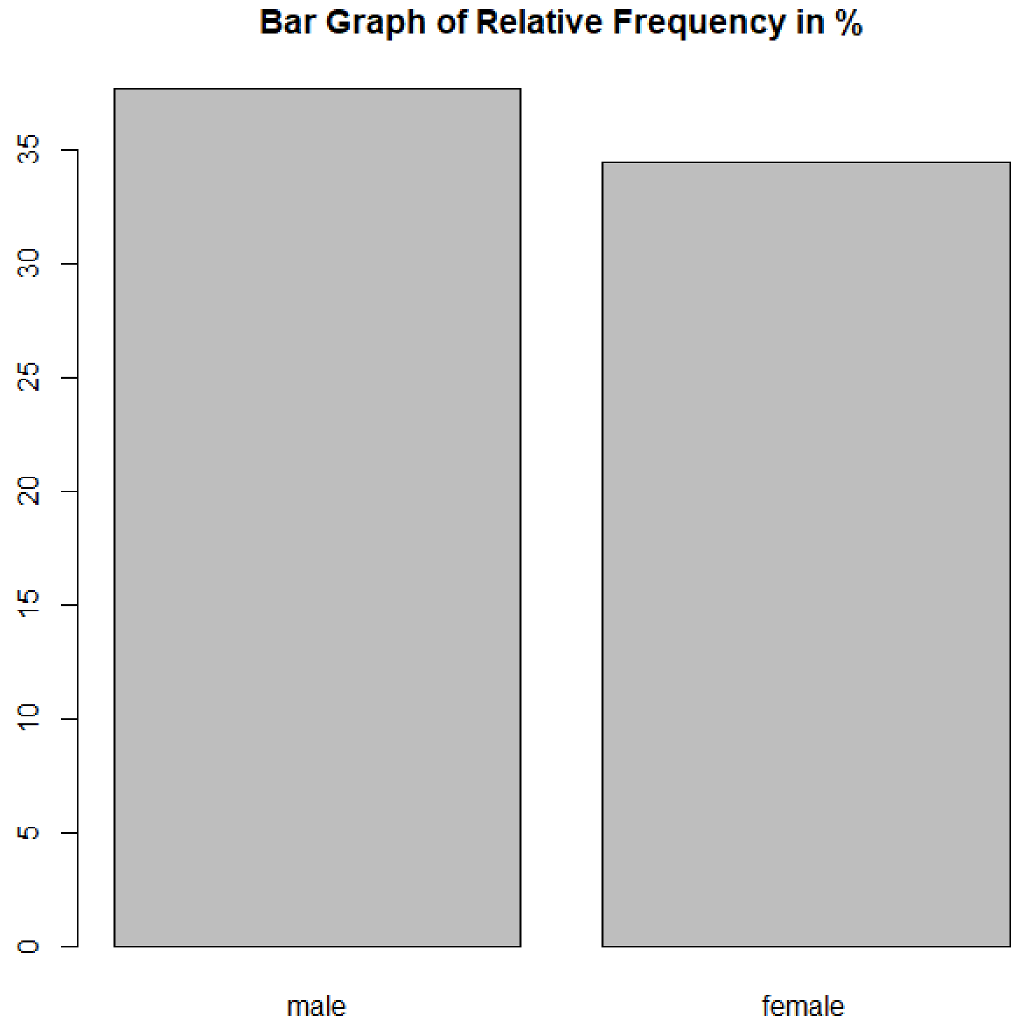
Example: The mean blood pressure among all adults aged 30 to 75 in the US population.

4.1 Dichotomous Variables

Example:

	n	Number on Treatment	Relative Frequency (%)
Males	1622	611	37.7
Females	1910	608	31.8
Total	3532	1219	34.5

Description



R Code:

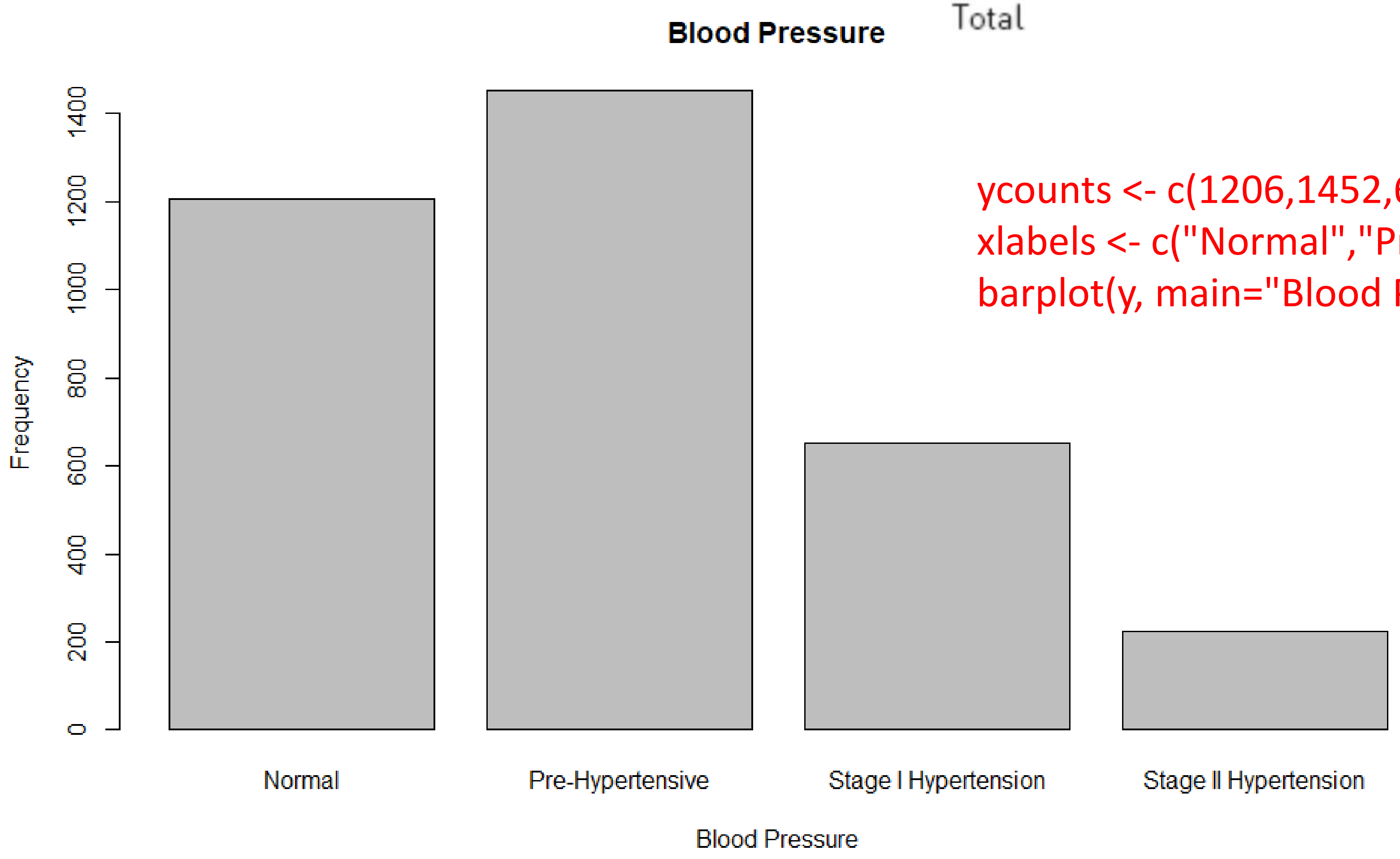
```
gender <- c("male", "female")
Relat_Freq <- c(37.7, 34.5)
barplot(Relat_Freq, names.arg=gender, main="Bar Graph of Relative Frequency in %")
```

4.2 Ordinal and Categorical Variables

Example:

	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative Relative Frequency (%)
Normal	1206	34.1	1206	34.1
Prehypertension	1452	41.1	2658	75.2
Stage I hypertension	653	18.5	3311	93.7
Stage II hypertension	222	6.3	3533	100.0
Total	3533	100.0		

Description




```
ycounts <- c(1206,1452,653,222)
xlabels <- c("Normal","Pre-Hypertensive","Stage I Hypertension", "Stage II Hypertension")
barplot(y, main="Blood Pressure",xlab="Blood Pressure", ylab="Frequency",names.arg=xlabels)
```

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Mean

Notation for sum x's

$$\bar{X} = \frac{\sum X}{n} = \frac{12}{5} = 2.4$$

$$\sum X = 1 + 2 + 2 + 3 + 4 = 12$$

Notation for sum x's

```
x <- c(1,2,2,3,4)
sum(x)
mean(x)
```

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Median

$$median = middle \quad value$$
$$median = 2$$

Order data from smallest to largest.
If the number of data values is odd,
take the middle value as the median.
If the number of data values is even,
take the average of the middle two.

```
x <- c(1,2,2,3,4)
median(x)
```

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Mode

$$mode = most \quad frequent \quad value$$
$$mode = 2$$

Order data from smallest to largest. Count how many time each value occurs. Take the one with the highest count.

```
get_mode <- function(x) {
  uniq_x <- unique(x)
  uniq_x[which.max(tabulate(match(x, uniq_x)))]
}

x <- c(1,2,2,3,4)
mode(x)
```

4.3 Continuous Variables

	X	$\bar{X}$	$X-\bar{X}$	$(X-\bar{X})^2$
	1	2.4	-1.4	1.96
	2	2.4	-0.4	0.16
	2	2.4	-0.4	0.16
	3	2.4	0.6	0.36
	4	2.4	1.6	2.56
$\sum$	12			5.20

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Variance

$$s^2 = \frac{1}{n-1}\sum(X-\bar{X})^2$$

$$s^2 = \frac{1}{5-1}\left[(1-2.4)^2+(2-2.4)^2+(2-2.4)^2+(3-2.4)^2+(4-2.4)^2\right]$$

$$s^2 = \frac{5.2}{4} = 1.3$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{1.3} = 1.14$$

```
x <- c(1,2,2,3,4)
var(x)
sd(x)
```

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,2,3,4

Sample Variance

X	X^2
1	1
2	4
3	9
3	9
4	16
$\sum$ 12	34

$$s^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{1}{n}\left(\sum X\right)^2\right]$$

$$s^2 = \frac{1}{5-1}\left[34 - \frac{12^2}{5}\right]$$

$$s^2 = \frac{5.2}{4} = 1.3$$

$$s = \sqrt{s^2} = \sqrt{1.3} = 1.14$$

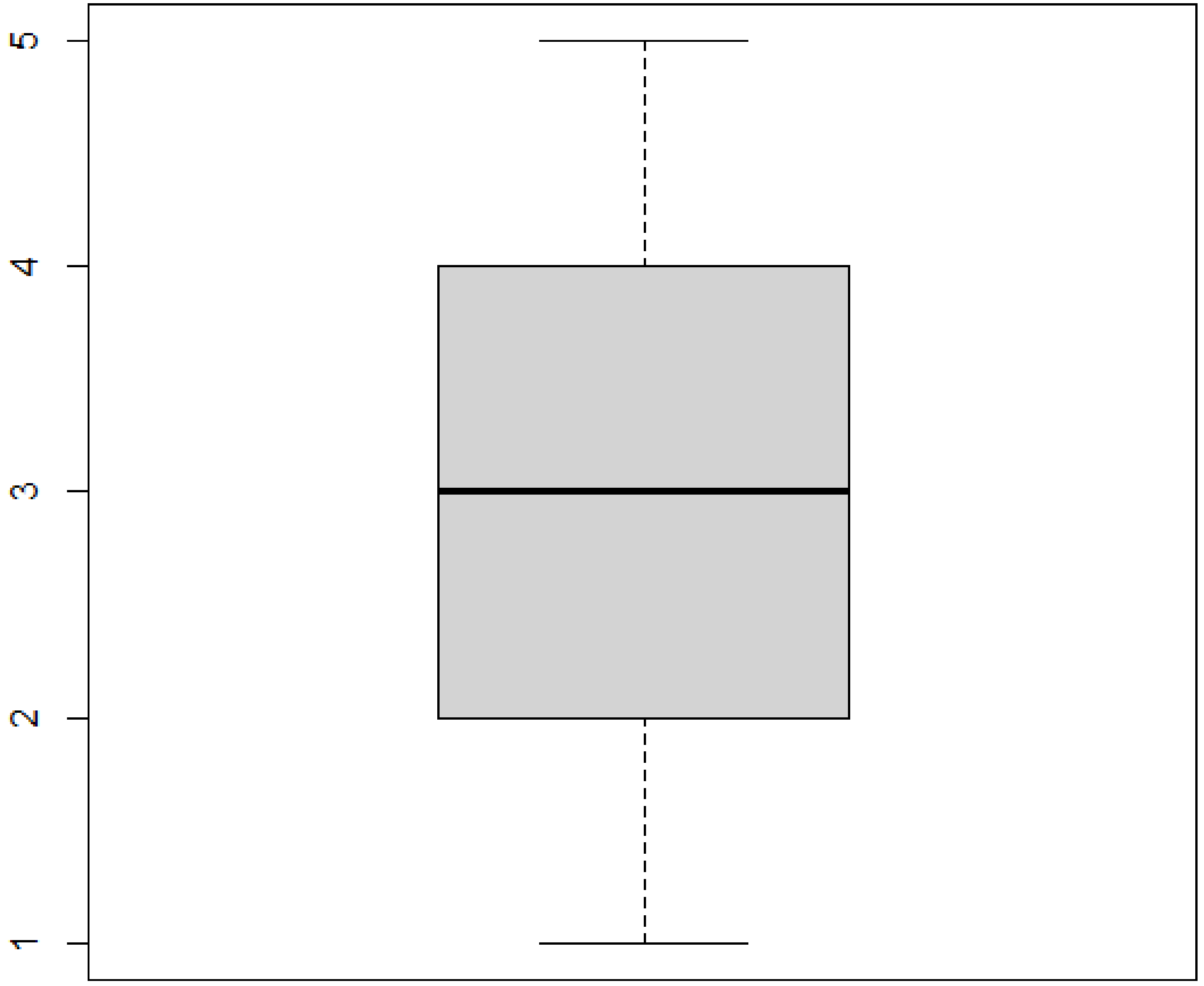
```
x <- c(1,2,2,3,4)
n <- length(x)
x2=x*x
sum(x)
sum(x2)
s2 <- (sum(x2)-sum(x)^2/n)/(n-1)
s2
s <- sqrt(s2)
s
```

4.3 Continuous Variables

Example 1: Small Numbers

Data values: 1,2,3,4,5

Box-Whisker Plot



5-number summary

1. L = minimum value
2. Q_1 = data value where 25% are smaller
3. Q_2 = median (where 50% are smaller)
4. Q_3 = data value where 75% are smaller
5. H = maximum value

```
x <- c(1,2,3,4,5)
boxplot(x)
quantile(x, probs = c(0,0.25,0.50,0.75,1))
```

```
0%  25%  50%  75% 100%
 1    2    3    4    5
```

4.3 Continuous Variables

Q_1 = data value where 25% are smaller
Q_3 = data value where 75% are

Inter Quartile Range

$$IQR = Q_3 - Q_1$$

Outliers

are below $Q_1 - 1.5\ IQR$

or above $Q_3 + 1.5\ IQR$

No outliers, use the mean and standard deviation to summarize the sample.

Outliers, use the median and IQR to summarize the sample.

4.3 Continuous Variables

Example 1: Diastolic Blood Pressures

Data values: 62,63,64,67,70,72,76,77,81,81

Sample Mean

Notation for sum x's

$$\bar{X} = \frac{\sum X}{n} = \frac{713}{10} = 71.3$$

$$\sum X = 62+63+64+67+70+72+76+77+81+81 = 713$$

Notation for sum x's

```
x <- c(62, 63, 64, 67, 70, 72, 76, 77, 81, 81)
sum(x)
mean(x)
```

4.3 Continuous Variables

Example 1: Diastolic Blood Pressures

Data values: 62,63,64,67,70,72,76,77,81,81

Sample Median

$$median = middle \quad value$$
$$median = 71$$

Order data from smallest to largest.
If the number of data values is odd,
take the middle value as the median.
If the number of data values is even,
take the average of the middle two.

```
x <- c(62, 63, 64, 67, 70, 72, 76, 77, 81, 81)
median(x)
```

4.3 Continuous Variables

Example 1: Diastolic Blood Pressures

Data values: 62,63,64,67,70,72,76,77,81,81

Sample Mode

$$mode = most \quad frequent \quad value$$
$$mode = 81$$

Order data from smallest to largest. Count how many time each value occurs. Take the one with the highest count.

4.3 Continuous Variables

Example 1: Diastolic Blood Pressures

Data values: 62,63,64,67,70,72,76,77,81,81

Sample Variance

$$s^2 = \frac{1}{n-1}\sum(X-\bar{X})^2$$

$$s^2 = \frac{1}{10-1}\left[(62-71.3)^2+\ldots+(81-71.3)^2\right]$$

$$s^2 = \frac{472.9}{9} = 52.5$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{52.5} = 7.24$$

```
x <- c(62, 63, 64, 67, 70, 72, 76, 77, 81, 81)
var(x)
Sd(x)
```

4.3 Continuous Variables

Example 1: Diastolic Blood Pressures

Data values: 62,63,64,67,70,72,76,77,81,81

Sample Variance

$$s^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{1}{n}\left(\sum X\right)^2\right]$$

$$s^2 = \frac{1}{10-1}\left[51309 - \frac{713^2}{10}\right]$$

$$s^2 = \frac{472.9}{9} = 52.5$$

$$s = \sqrt{s^2} = \sqrt{52.5} = 7.24$$

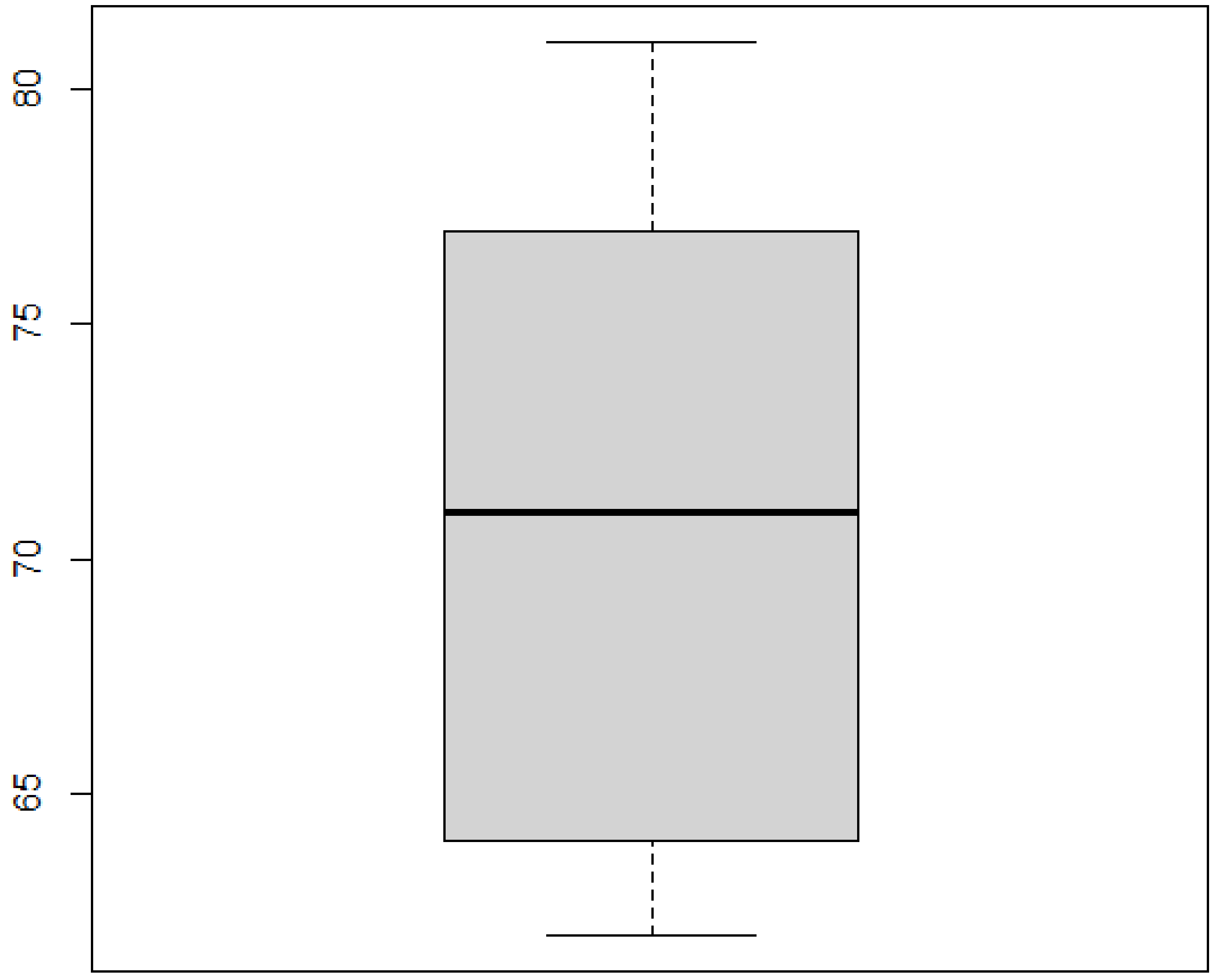
```
x <- c(62, 63, 64, 67, 70, 72, 76, 77, 81, 81)
n <- length(x)
x2=x*x
sum(x)
sum(x2)
s2 <- (sum(x2)-sum(x)^2/n)/(n-1)
S2
s <- sqrt(s2)
s
```

4.3 Continuous Variables

Example 1: Diastolic Blood Pressures

Data values: 62,63,64,67,70,72,76,77,81

Box-Whisker Plot



5-number summary

1. L = minimum value
2. Q_1 = data value where 25% are smaller
3. Q_2 = median (where 50% are smaller)
4. Q_3 = data value where 75% are smaller
5. H = maximum value

$$IQR = Q_3 - Q_1$$

```
x <- c(62, 63, 64, 67, 70, 72, 76, 77, 81, 81)
boxplot(x)
quantile(x, probs = c(0,0.25,0.50,0.75,1))
```

```
   0%   25%   50%   75%  100%
62.00 64.75 71.00 76.75 81.00
```

Questions?

Homework 4

Read Chapter 4.

Problems # 2, 4, 6, 7, 9