Section 1.5
Graphs and Describing Distributions

Data can be displayed using graphs. Some of the most common graphs used in statistics are:

- Bar graph
- Pie Chart
- Dot plot
- Histogram
- Stem and leaf plot
- Box plot
- Cumulative Frequency plot

The question is, what type of graph would be best for our data?

For categorical data, bar graphs and pie charts are the best representations.

A **bar graph** used to compare the sizes of different groups. They are created by listing the categorical data along the *x*-axis and the frequencies along the *y*-axis. Bars are drawn above each categorical variable and the height of each bar indicates the size of the respective group. R command: barplot( )

A **pie chart** is a circular chart, divided into sectors, indicating the proportion of each data value compared to the entire set of values. To create one, percentage values for each category must be calculated first. R command: pie( )

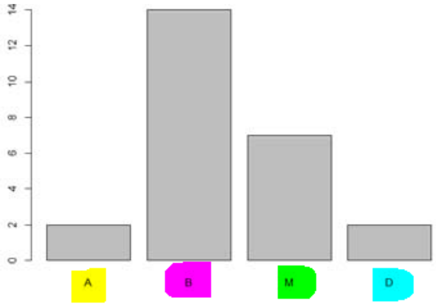
Example 1: Twenty-five students were asked on a survey to give the highest educational degree of their parents. The results are given in the table below.

Education Level	**Number of Students**
Associate’s Degree	2
Bachelor’s Degree	14
Master’s Degree	7
Doctorate degree	2
Total	***25***

Create a bar graph.
Commands:

```
degree=c(2,14,7,2)
barplot(degree, names.arg=c(“A”,”B”,”M”,”D”))
```

Result:



Dot Plot, Histogram, Stem and Leaf Plot, and Box Plot

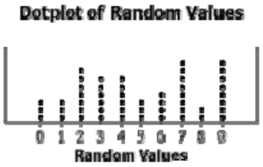
Quantitative data can be displayed using any of the other types of graphs. Which one we choose usually depends on how much data we have.

Example 2: Height measurements for a group of people were taken. The results are recorded below (in inches):

66, 68, 63, 71, 68, 69, 65, 70, 73, 67, 62, 59, 63, 68, 71, 63, 63, 60, 64, 66, 58

We will organize this data using different graphs:

A **dot plot** is a graph consisting of points plotted on simple scale. To create one, simply place dots above the values listed on a number line. The number of dots above each value is determined by how many times that respective value appears in the data set.
R command: stripchart( )



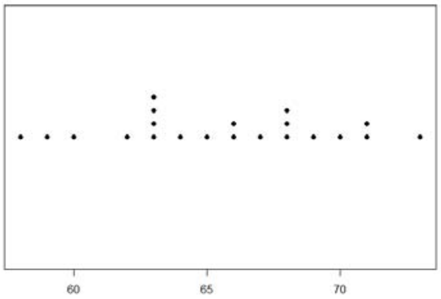
a. Create a dot plot for the data. *If doing by hand, order the data first.*
Commands:

```
heights=c(66,68,63,71,68,69,65,70,73,67,62,59,63,68,71,63,63,60,64,66,58)
stripchart(heights,method=”stack”,pch=16,offset=1)
```

filled dots

spacing dots

Result:



A **stem and leaf plot** shows exact values of individual observations. The digits in the largest place are referred to as the stem and the digits in the smallest place are referred to as the leaf (leaves). The leaves are displayed to the right of the stem.
R command: stem( )

```
1 | 2 3 4 5 5 9
2 | 0 2 2 4 6 8
3 | 1 1 2 3
```

b. Create a stem and leaf plot for the data. *If doing by hand, order the data first.*

Command:

```
stem(heights)
```

length(heights) → 21

Result:

```
5 | 89
6 | 0233334
6 | 56678889
7 | 0113
```

What is the median of this stem leaf plot?

10 1 10
med

$\boxed{66}$

$$\frac{21-1}{2} = 10$$

Try another: What is the median of stem leaf plot?

```
1 | 2 3 4 5 5 9
2 | 0 2 2 4 6 8
3 | 1 1 2 3
```

16 values

7 2 7

$$\frac{16-2}{2} = 7$$

$$\frac{22+22}{2} = \boxed{22}$$

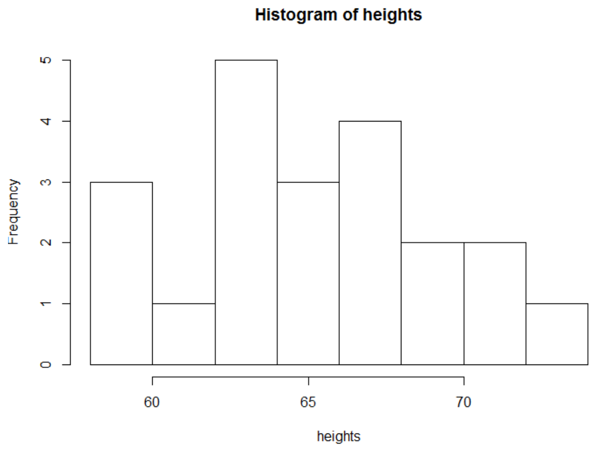
Histograms show how observations are distributed across groups, but do not show the exact values of individual observations. To create one, first divide the data into classes, or bins, of equal width. Next, count the number of observations in each class. The horizontal axis will represent the variable values and the vertical axis will represent your frequency or your relative frequency.
R command: hist( )

c. Create a histogram for the data.

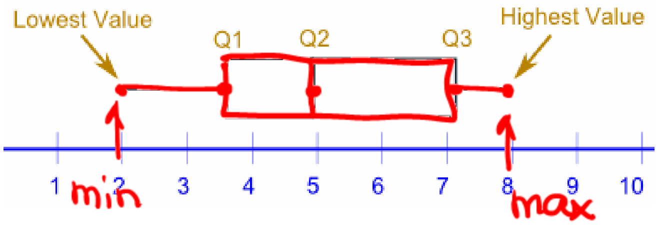
Command:

```
hist(heights)
```

Result:



Boxplots or sometimes called **Box and Whisker Plots** not only help identify features about our data quickly (such as spread and location of center) but can be very helpful when comparing data sets. They display patterns of quantitative data.

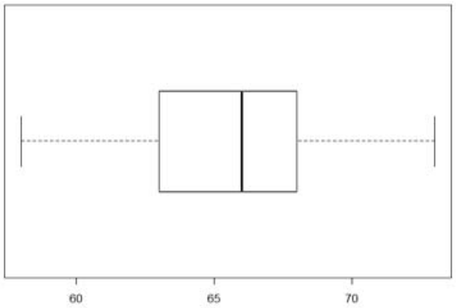


To create one, find the five number summary. Plot each number and draw a box starting at Q1 to Q3. Then draw a line segment within the box to represent the median. Connect the min and max to the box with line segments. This is the boxplot. If data contains outliers, a **box and whiskers plot** can be used instead to display the data. In a box and whiskers plot, the outliers are displayed with dots above the value and the segments begin (or end) at the next data value within the outlier interval.

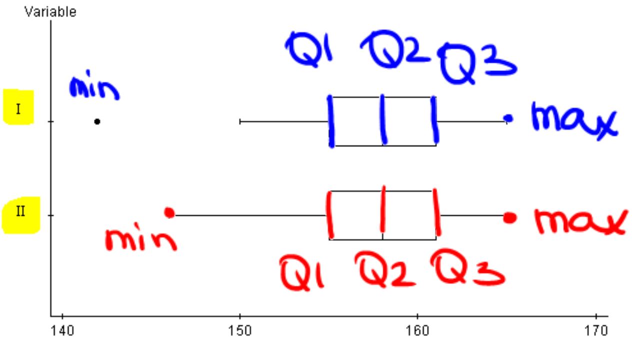
R command: boxplot( )

d. Create a boxplot for the data.
Command:
boxplot(heights,horizontal=TRUE)

Result:



Example 3: Use the following boxplots to determine which of the following statements about these data sets CANNOT be justified?



a. The interquartile range of data set I is equal to the interquartile range of data set II.

IQR = Q3 - Q1 YES

b. The range of data set I is greater than the range of data set II.

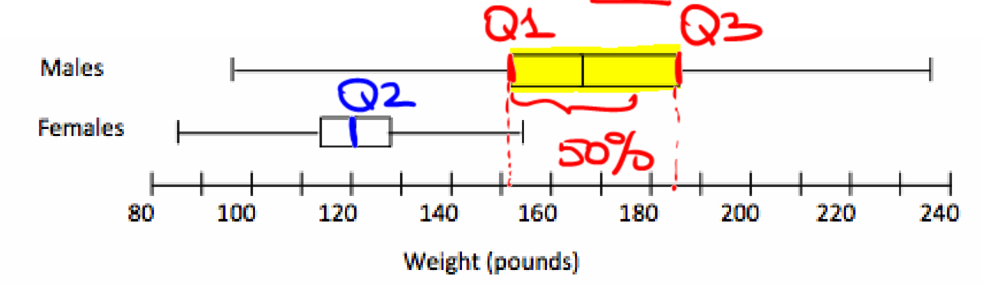
Range = max - min YES

c. Data set I and data set II have the same number of data points. NO

Cannot be justified.

Box plots and Histograms do not show exact values.

Example 4: The weights of male and female students in a class are summarized in the following boxplots. Which of the following is NOT correct?



a. The male students have less variability than the female students.

False

b. About 50% of the male students have weights between 150 and 185 lbs.

True

c. The mean weight of the female students is about 120 because of symmetry.

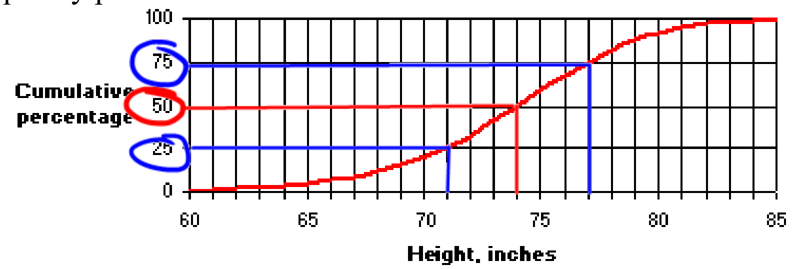
True

Q2 = median

Cumulative Frequency Plot

Sometimes we are interested in the relative position of observation values, i.e. finding the percentile from a graph. A **cumulative frequency plot** of the percentages (also called an **ogive**) can be used to view the total number of events that occurred up to a certain value. Percentiles can be easily viewed using this type of graph.

Example 5: Heights (in inches) of a set of basketball players are indicated below in the cumulative frequency plot.



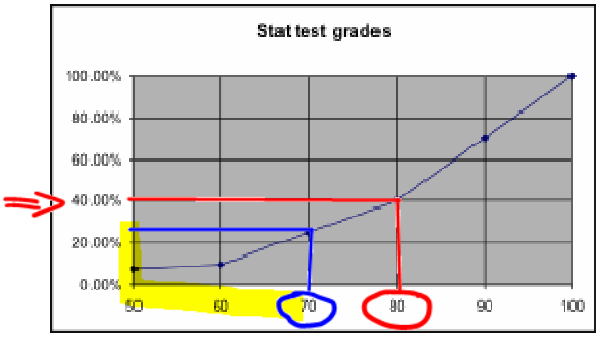
a. What is the median?

about 74 inches

b. What is the interquartile range?

IQR = Q3 - Q1 = 77 - 71 = 6 inches

Example 6: The figure below shows a cumulative relative frequency plot of 40 scores on a test given in a Statistics class. Which of the following conclusions can be made from the graph?



a. If the passing score is 70, most students did not pass the test.

False. Only about 25% of students have a score less than 70.

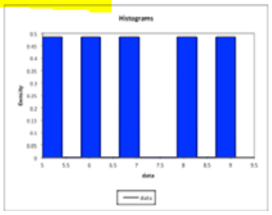
b. Sixty percent of the students had a test score above 80.

True

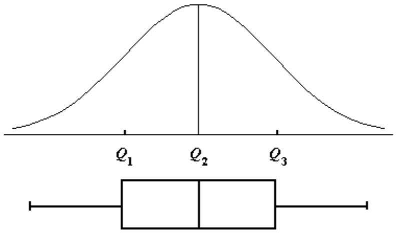
It is very helpful in statistics to understand the summary features of quantitative variables by looking at their graphs. Center, spread and shape are the most important features to note. There are many patterns that our distribution can follow.

Patterns and Shapes

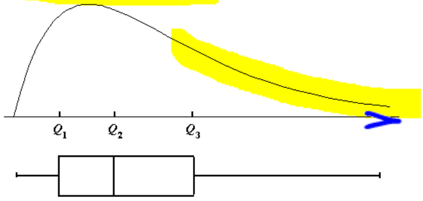
Uniform – all data has the same height on the graph.



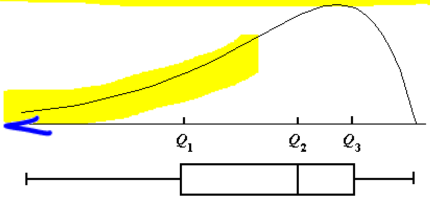
Symmetric – same shape to the left and right of the center. The first graph is a bell shaped curve. The mean and median will be equal in a bell shaped distribution.



Skewed Right (or positive skew) – longer tail on the right side. The mean will be larger than the median in a skewed right distribution.



Skewed Left (or negative skew) – longer tail on the left side. The mean will be smaller than the median in a skewed left distribution.



Example 7: Given the following stem and leaf plot

1	0 2 2 3 5 7 8 9
2	1 1 2 3 4
3	2
4	
5	9

a. Is the data skewed to the left or to the right? What can be said about its mean?

Right mean > median

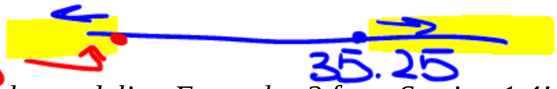
b. Any outliers?

Possibly 59 (needs to be checked!)

YES! $Q3+1.5(IQR) = 35.25$

$59 > 35.25$

$Q1-1.5(IQR)$



We can verify this by modeling Examples 2 from Section 1.4!

data= c ("enter data")

	min	Q1	Q2	Q3	max
fivenum (data) →	10	14	19	22.5	59

$IQR = Q3-Q1 = 22.5-14 = 8.5$

$$Q3+1.5(IQR) = 22.5 \overset{②}{+} \overset{①}{1.5*8.5} = 35.25$$