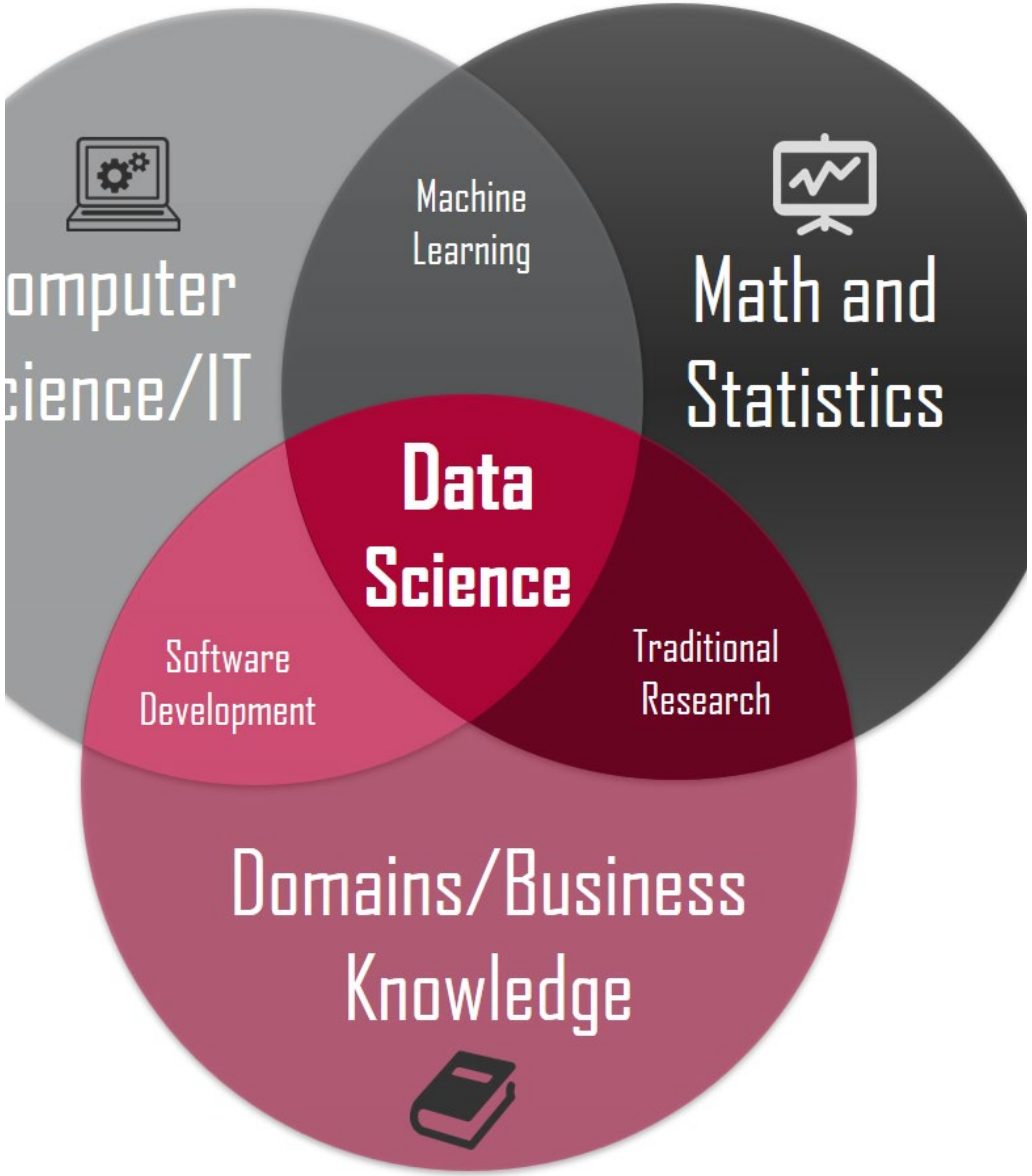


https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

THE UNIVERSITY OF WESTERN AUSTRALIA

STAT 1400

Statistics for Science

Lecture Week 3

Dr Darfiana Nur

Department Mathematics and Statistics

Aims of this week

- Aim 1 Numerical Summaries
  - 3S - Shape; (S)Center; Spread
- Aim 2 The Boxplot
- Aim 3 Data Screening and Outliers

EXAMINING RELATIONSHIPS, 2 VARIABLES

- Aim 4 Explanatory and response variables
- Aim 5 Relationship Between a Continuous Response Variable (Y) and a Categorical Explanatory (X) Variable
- Aim 6 Relationship Between Two Categorical Variables (UWA Sem 1 2023)

References: Moore et al (2021) Chapter 1

Aim 1 Numerical summaries for *numerical* data

THE UNIVERSITY OF WESTERN AUSTRALIA

- In Statistics we often use
  summary statistics and
  graphs to represent samples of data

- This allows us to efficiently present information and provides a basis for comparison and tentative conclusions

Numerical summaries for *numerical* data

- Numerical summaries (statistics) for **‘center’** or **location**
    1. mode
    2. median
    3. mean
- Numerical summaries (statistics) for **spread**
    1. **range**
    2. **inter-quartile range (IQR)**
    3. **standard deviation**

Measure of center / location 1: Mode

- The value of the variable that occurs most frequently.

In-class Exercise 1

Data: 7, 2, 5, 1, 5, 5, 3, 2, 12

Mode = ?

Measure of center /location 2: Median

- **Median**

middle value of data set

divides *ordered* data into two equal parts

Example 2 odd number of data points

Data values : 5, 9, 4, 8, 7

Ordered data : 4, 5, **7**, 8, 9

median

Median (continued)
Example 3 even number of data points

Data values: 5, 9, 4, 8, 7, 9

Order the data: 4, 5, 7, 8, 9, 9

When there is an even number of data points the median is the average of the middle two

i.e. median = (7+ 8)/2 = 7.5

Measure of center 2: the median

The **median** is the midpoint of a distribution—the number such that half of the observations are smaller and half are larger.

Example 4: Years until death for a certain disease

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9
12	12	3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6
25	**12**	**6.1**

1. Sort observations by size.
n = number of observations

2.a. If n is **odd,** the median is observation $(n+1)/2$ down the list

← $n = 25$
$(n+1)/2 = 26/2 = 13$
Median = 3.4

$n = 24$ →
$n/2 = 12$
Median = (3.3+3.4) /2 = 3.35

2.b. If n is **even,** the median is the mean of the two middle observations.

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9
12		3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6

Measure of center 3: the mean
Example 5 Women's height

The mean or arithmetic average

To calculate the *average,* or **mean,** add all values, then divide by the number of cases. It is the "center of mass."

Sum of heights is 1598.3
divided by 25 women = 63.9 inches

58.2	64.0
59.5	64.5
60.7	64.1
60.9	64.8
61.9	65.2
61.9	65.7
62.2	66.2
62.2	66.7
62.4	67.1
62.9	67.8
63.9	68.9
63.1	69.6
63.9	

In-Class Exercise 2. What is the median?

woman (i)	height (x)		woman (i)	height (x)
i = 1	x_1= 58.2		i = 14	x_{14}= 64.0
i = 2	x_2= 59.5		i = 15	x_{15}= 64.5
i = 3	x_3= 60.7		i = 16	x_{16}= 64.1
i = 4	x_4= 60.9		i = 17	x_{17}= 64.8
i = 5	x_5= 61.9		i = 18	x_{18}= 65.2
i = 6	x_6= 61.9		i = 19	x_{19}= 65.7
i = 7	x_7= 62.2		i = 20	x_{20}= 66.2
i = 8	x_8= 62.2		i = 21	x_{21}= 66.7
i = 9	x_9= 62.4		i = 22	x_{22}= 67.1
i = 10	x_{10}= 62.9		i = 23	x_{23}= 67.8
i = 11	x_{11}= 63.9		i = 24	x_{24}= 68.9
i = 12	x_{12}= 63.1		i = 25	x_{25}= 69.6
i = 13	x_{13}= 63.9		n=25	Σ=1598.3

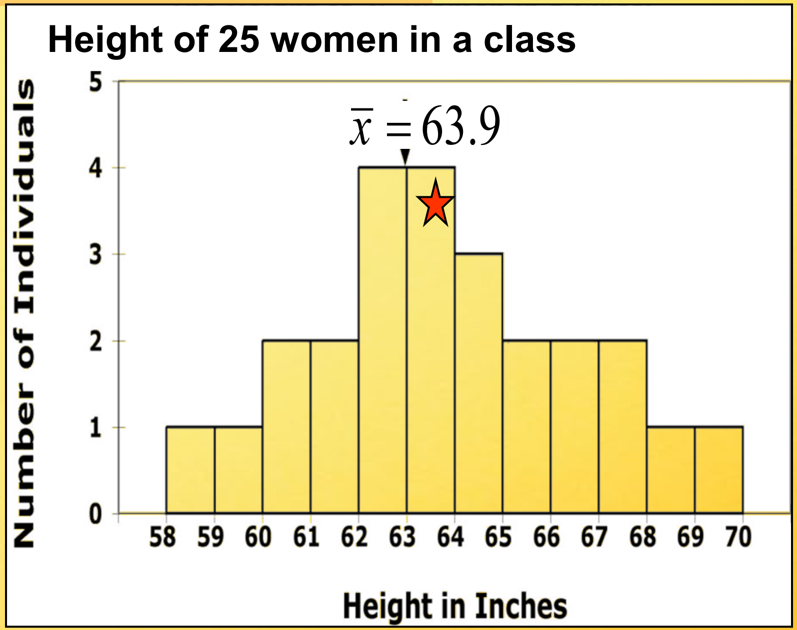
Mathematical notation:

Data x_i , i=1,2, …, n

Sample mean $\bar{x}$

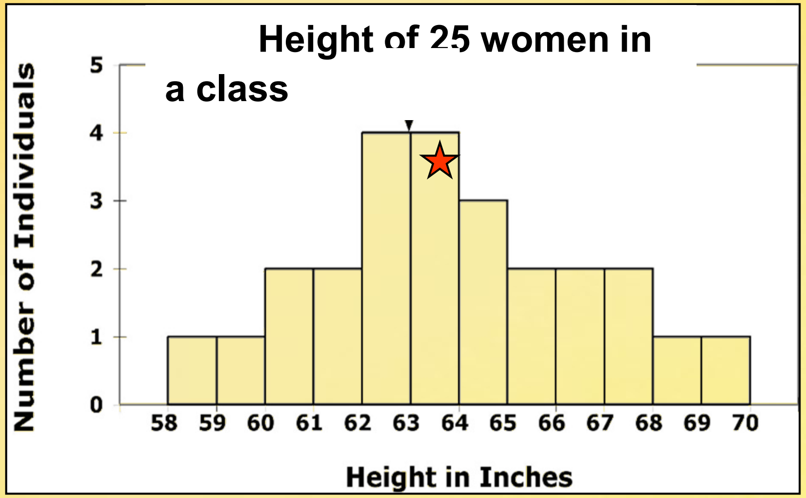
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1598.3}{25} = 63.9$$



Learn right away how to get the mean using calculator.

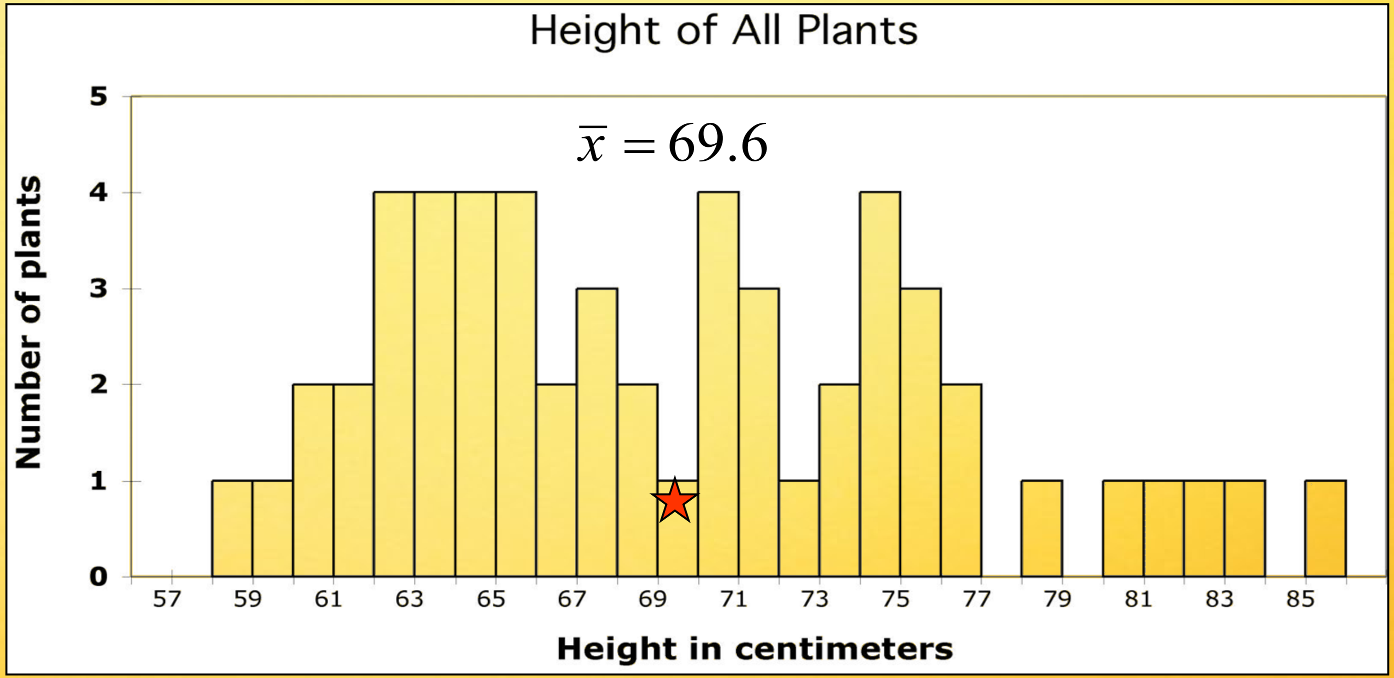
Your numerical summary must be meaningful

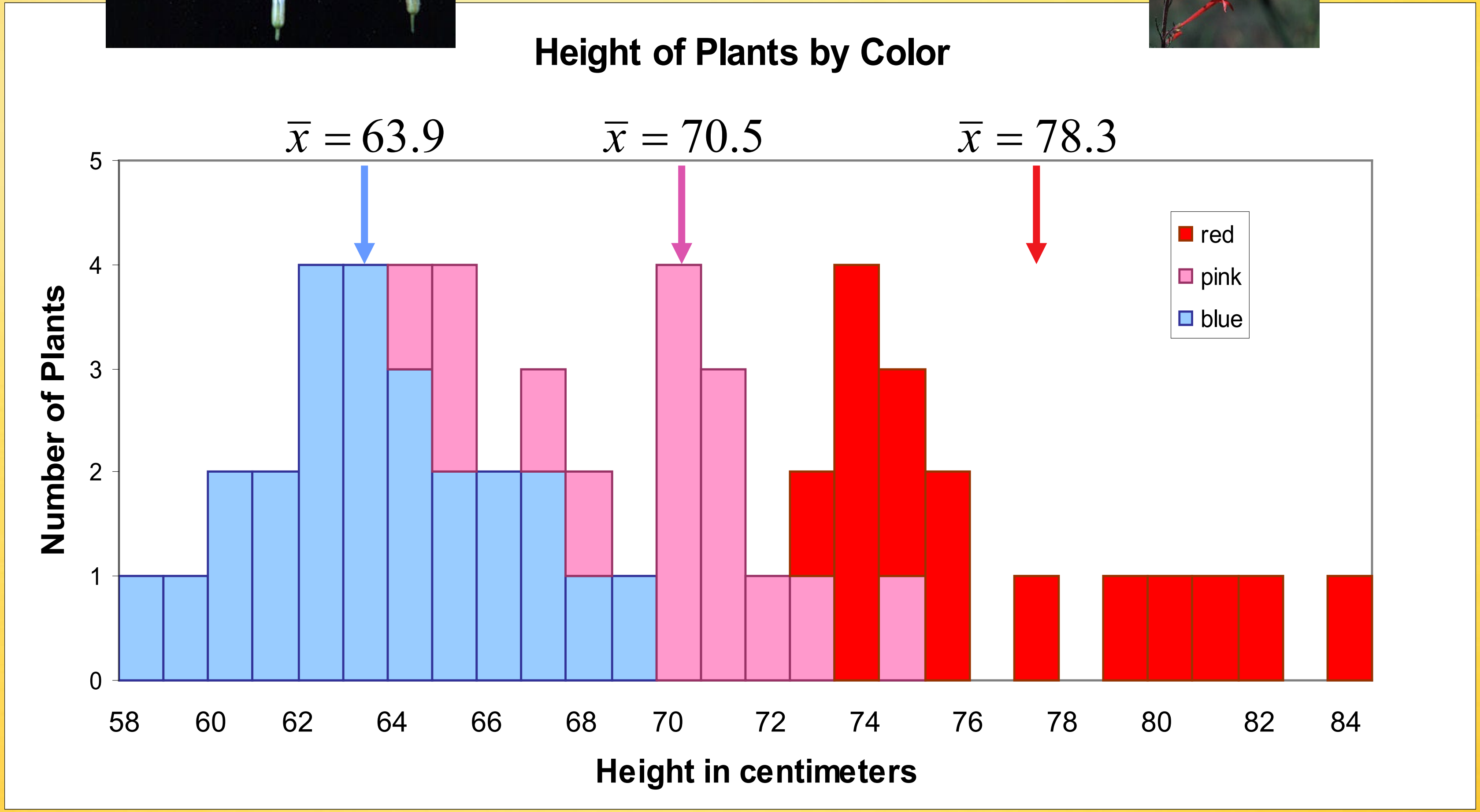


The distribution of women's heights appears coherent and symmetrical. The mean is a good numerical summary.

Example 6: Height of Plants

Here the shape of the distribution is wildly irregular.
Why?
Could we have more than one plant species or phenotype?

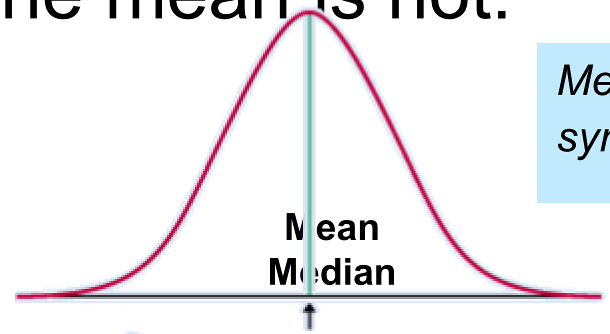




A single numerical summary here would not make sense.

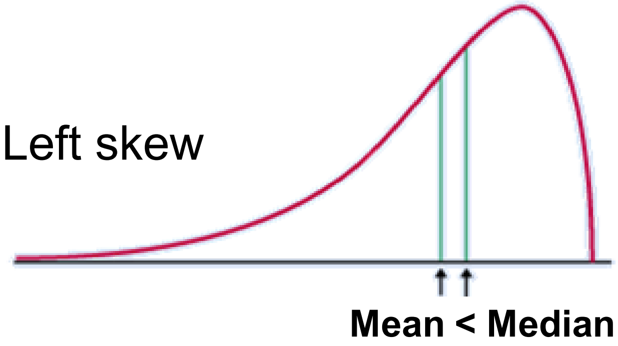
Comparing the mean and the median

- The mean and the median are the same only if the distribution is symmetrical.
- The median is a measure of center that is resistant or robust to skew and outliers. The mean is not.

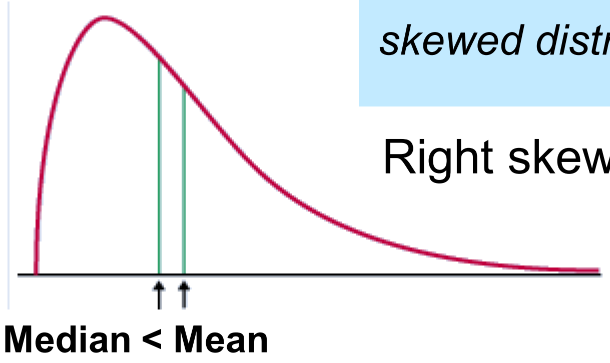


Mean and median for a symmetric distribution

Left skew



Mean and median for skewed distributions



Right skew

Comparison between mean and median: Which one is better?

Both are useful for indicating the center of a data set.

Mean is more commonly used but is affected by extreme values (outliers) and skewness

Median may be a better representation of the 'typical' value for skewed data OR data with extreme values because the sample is split in half.

Measure of Spread 1: Range

- Range is the difference between largest (maximum) and smallest (minimum) values in the data set.
- Sensitive to unusually extreme values (i.e., values at the ends of distribution)

In-class Exercise 3

Data values:

21, 25, 23, 28, 16, 19, 17, 21, 15, 22

maximum = ?   minimum = ?

range = ?

Quartiles

THE UNIVERSITY OF WESTERN AUSTRALIA

- First 25% of data are less than first quartile Q1 (and 75% of data are greater than Q1)
- Second quartile Q2 is the median, with 50% of data on either side
- First 75% of data are below the third quartile Q3 (and 25% of data are greater than Q3)

The quartiles

Example 7 Years until death for a certain disease

The **first quartile, Q_1**, is the value in the sample that has 25% of the data at or below it (it is the median of the lower half of the sorted data, excluding *M*).

The **third quartile, Q_3**, is the value in the sample that has 75% of the data at or below it (it is the median of the upper half of the sorted data, excluding *M*).

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	1	2.3
9	2	2.5
10	3	2.8
11	4	2.9
12	5	3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	1	4.7
22	2	4.9
23	3	5.3
24	4	5.6
25	5	6.1

Q_1= first quartile
=(2.1+2.3)/2== 2.2

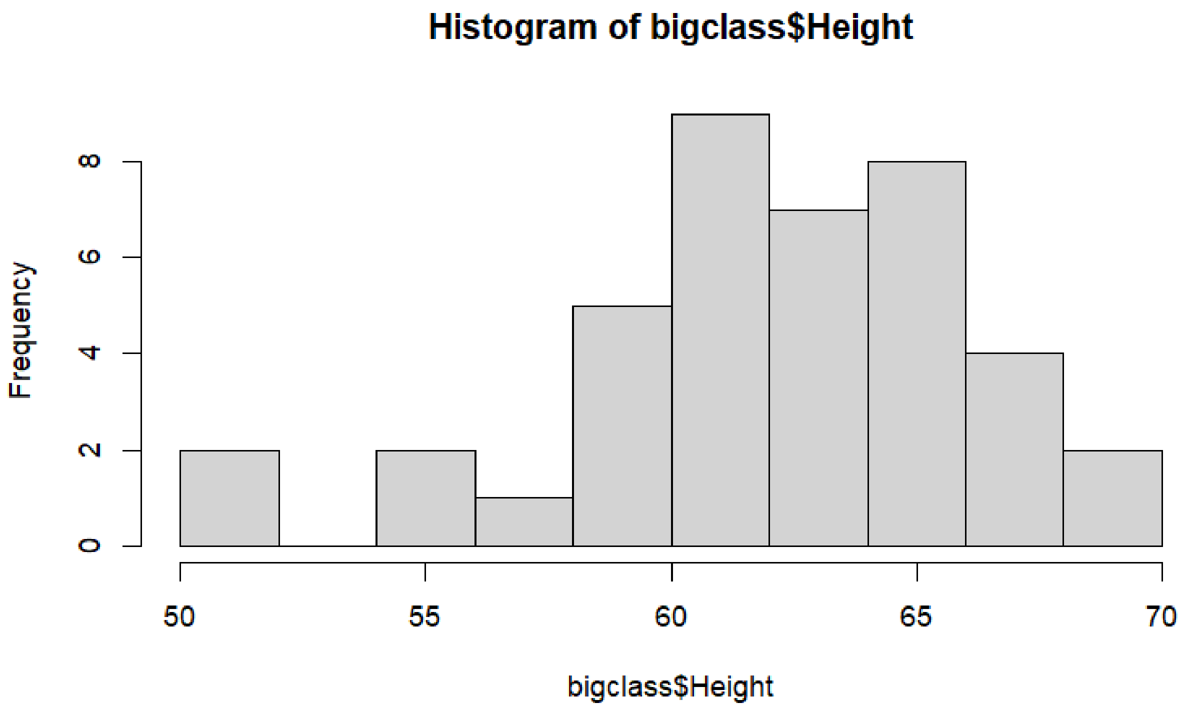
M = median = *3.4*

Q_3= third quartile
=(4.2+4.5)/2= = 4.35

Finding the Quartiles is not difficult but can be tedious so we will generally use R

(*for "height" in Big Class data set (Computer Lab 2)*)

````
```{r}
summary(bigclass$Height)
```
````



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
51.00	60.75	63.00	62.55	65.00	70.00

Measure of spread 2: Inter-Quartile Range (IQR)

- The IQR is the difference between Q1 and Q3:

IQR=Q3-Q1

```{r}
IQR(bigclass$Height)
```

- For the previous example

  Q1 = ? and Q3 = ?

  IQR = ?

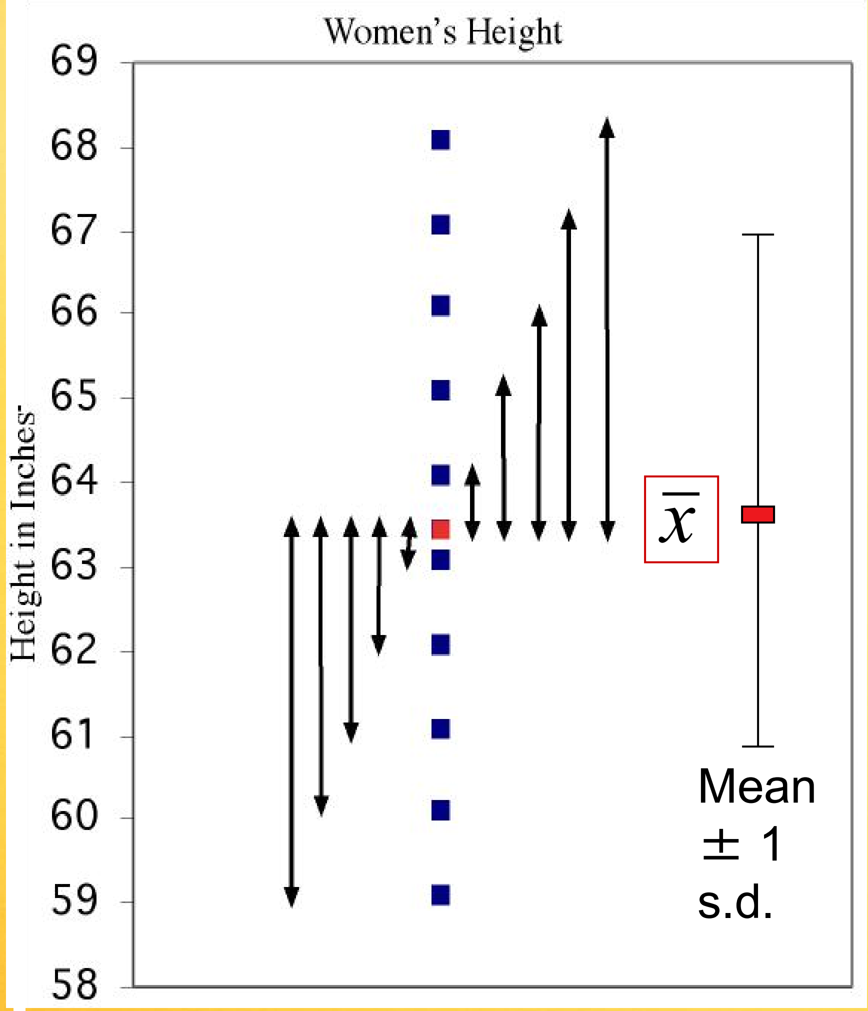
  IQR measures the spread of the middle 50% of the data.

- It is not sensitive to extreme values. Why?

Measure of spread 3: the standard deviation

Example 8 Women's height

The standard deviation "s" is used to describe the variation around the mean. Like the mean, it is not resistant to skew or outliers.



1. First calculate the **variance s^2.**

$$s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2$$

Data X_i , i=1,2, …, n

Sample mean $\bar{x}$

- n: sample size
- $\sum$: sum of

2. Then take the square root to get the **standard deviation s.**

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2}$$

Calculations …

$$s = \sqrt{\frac{1}{df}\sum_{1}^{n}(x_i - \bar{x})^2}$$

Mean = $\bar{x}$ = 63.4 n=14

Sum of squared deviations from mean = 85.2

Degrees freedom (df) = (n − 1) = 14-1=13

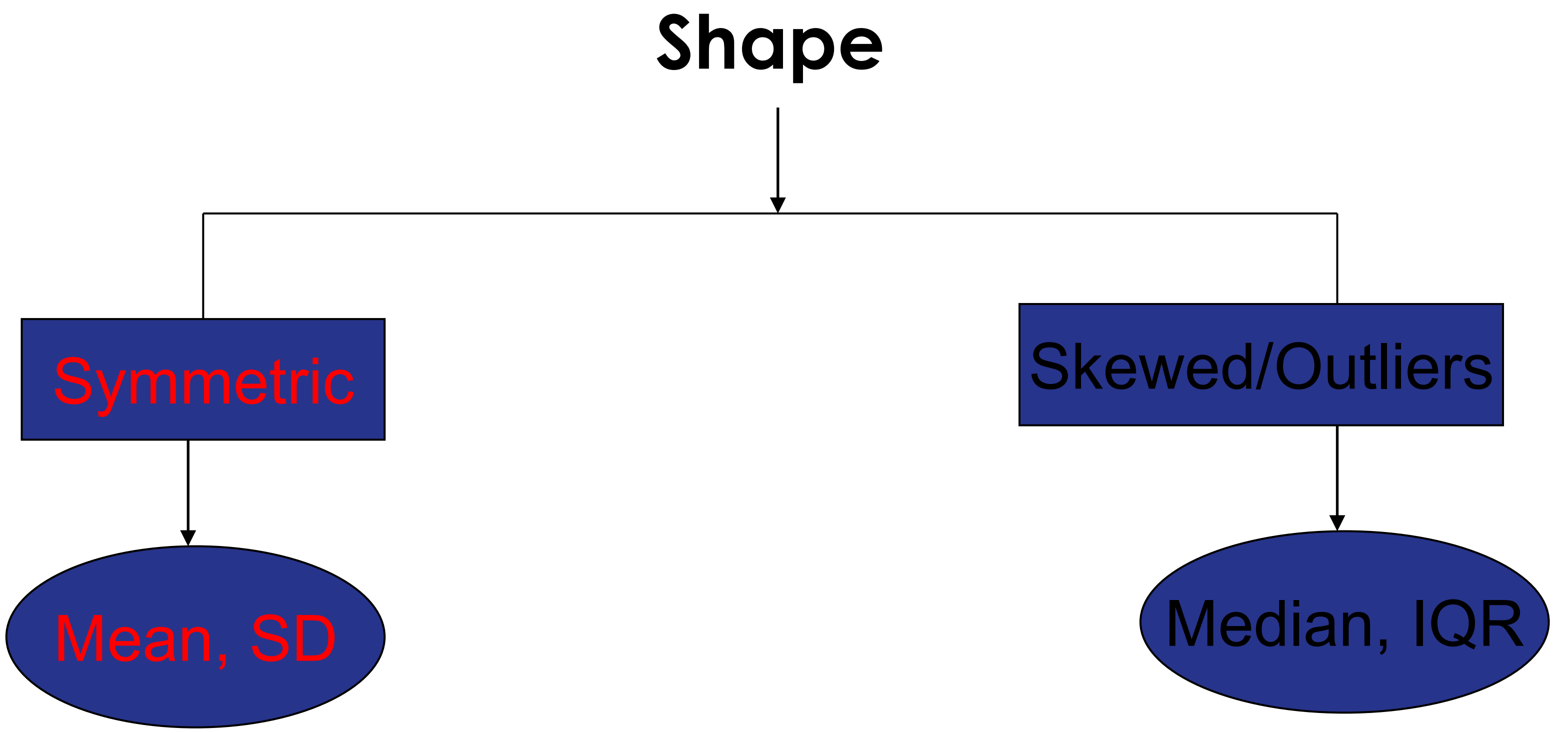
s^2 = variance = 85.2/13 = 6.55 inches squared

s = standard deviation = √6.55 = 2.56 inches

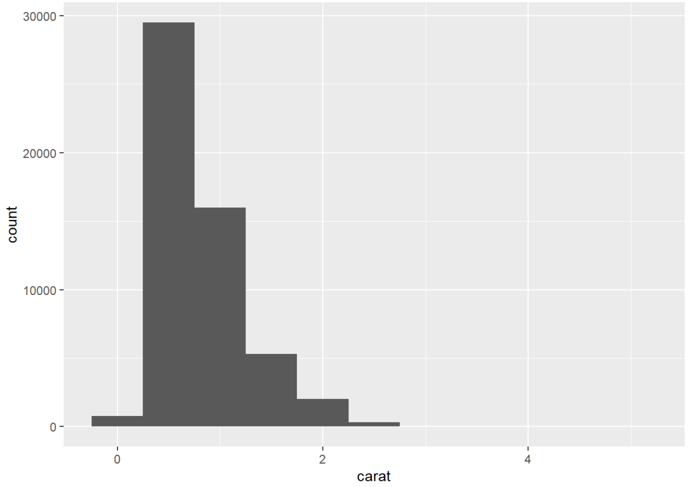
Women's height (inches)

i	x_i	$\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$
1	59	63.4	-4.4	19.0
2	60	63.4	-3.4	11.3
3	61	63.4	-2.4	5.6
4	62	63.4	-1.4	1.8
5	62	63.4	-1.4	1.8
6	63	63.4	-0.4	0.1
7	63	63.4	-0.4	0.1
8	63	63.4	-0.4	0.1
9	64	63.4	0.6	0.4
10	64	63.4	0.6	0.4
11	65	63.4	1.6	2.7
12	66	63.4	2.6	7.0
13	67	63.4	3.6	13.3
14	68	63.4	4.6	21.6
	Mean 63.4		Sum 0.0	Sum 85.2

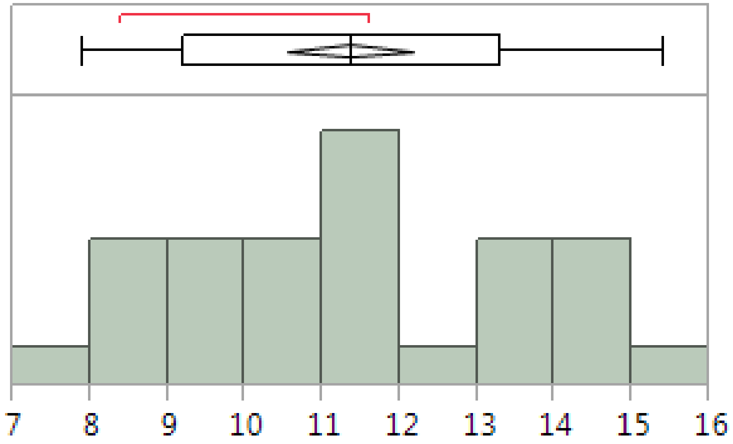
We'll rarely calculate these by hand, so make sure to know how to get the standard deviation using your calculator or R.

```
```{r}
var(bigclass$Height)
sd(bigclass$Height)
```
```

Variance and Standard Deviation

- ***Why do we square the deviations?***

The sum of the squared deviations of any set of observations from their mean is the smallest that the sum of squared deviations from any number can possibly be.

The sum of the deviations of any set of observations from their mean is always zero.

- ***Why do we emphasize the standard deviation rather than the variance?***

s, not s^2, is the natural measure of spread for Normal distributions.

s has the same unit of measurement as the original observations.

- ***Why do we average by dividing by n − 1 rather than n in calculating the variance?***

The sum of the deviations is always zero, so only $n - 1$ of the squared deviations can vary freely.

The number $n - 1$ is called the **degrees of freedom**.

Properties of Standard Deviation

- s measures spread about the mean and should be used only when the mean is the measure of center.
- $s = 0$ only when all observations have the same value and there is no spread. Otherwise, $s > 0$.
- s is not resistant to outliers.
- s has the same units of measurement as the original observations.

Interpreting measure of spread

- **Small** standard deviation implies the data is **concentrated around the mean.**

- **Large** standard deviation implies the data is **widely spread around the mean.**

- Can examine spread of data using histograms or box plots.

Comparison between IQR and SD

- Both are useful for indicating the spread of a data set.
- SD is more commonly used but is affected by outliers (ie. SD is sensitive to outliers)
- IQR is the best measure of spread for skewed data or data with extreme values because outliers have little effect on the IQR (ie IQR is insensitive to outliers)

Rule of thumb for choosing between Moment-based and Quantile-based measures



R output for summary statistics: Example 9

A data frame with 53940 rows and 10 variables:

- *price*: price in US dollars (\$326–\$18,823)
- *carat*: weight of the diamond (0.2–5.01)
- cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- color: diamond colour, from D (best) to J (worst)
- clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- x: length in mm (0–10.74); y: width in mm (0–58.9); z: depth in mm (0–31.8)
- depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
- table: width of top of diamond relative to widest point (43–95)

```
```{r}
summary(diamonds$carat)
```
```

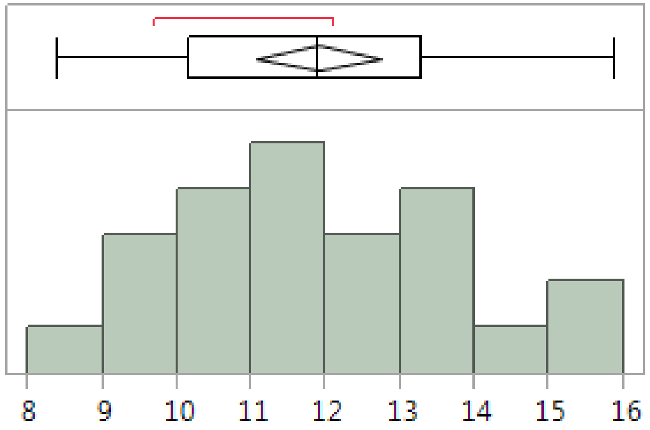
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2000	0.4000	0.7000	0.7979	1.0400	5.0100



Example 10: SCE Rates



Control/Normal



Smoker

sce.rate				
	Smoker	**Normal**		
Mean	11.89173913	11.34866667	Q1(Smoker)	10.435
Standard Error	0.408553629	0.395664938	Q3(Smoker)	13.225
Median	11.9	11.4	IQR(Smoker)	2.79
Mode	#N/A	13.3		
Standard Deviation	1.959354375	2.167146118		
Sample Variance	3.839069565	4.696522299	Q1(Normal)	9.33
Kurtosis	-0.386441966	-1.144451927	Q3(Normal)	13.3
Skewness	0.303634088	0.123651737	IQR(Normal)	3.97
Range	7.5	7.5		
Minimum	8.4	7.9		
Maximum	15.9	15.4		
Sum	273.51	340.46		
Count	23	30		

Changing the unit of measurement

- Variables can be recorded in different units of measurement. Most often, one measurement unit is a **linear transformation** of another measurement unit:

$$x_{new} = a + bx.$$

- Temperatures can be expressed in degrees Fahrenheit or degrees Celsius.

$$\text{Temperature}^{\text{Fahrenheit}} = 32 + (9/5)^{*}\ \text{Temperature}^{\text{Celsius}} \rightarrow a + bx.$$

Changing the unit of measurement

Linear transformations do not change the basic shape of a distribution (skew, symmetry, multimodal).

But they do change the measures of center and spread:

Multiplying each observation by a positive number *b* multiplies both measures of center (mean, median) and spread (IQR, *s*) by *b*.

Adding the same number *a* (positive or negative) to each observation adds *a* to measures of center and to quartiles but it does not change measures of spread (IQR, *s*).

Example 11: Forensic Science

A crime is committed at the city. A few hairs are found at the scene. The hair is analysed in a laboratory and a particular hair conditioner additive, distearoylethyl hydroxyethylmonium (DH), is found. The viscosity of DH (in cP) is considered an important clue to the type of conditioner used in the hair. Twelve measurements of DH taken from the hair at the crime scene are given below:

146 154 141 140 136 132 147 140 147 139 140 140

The summary statistics

Sample Mean	Sample Variance	Standard deviation	Median	Minimum	Maximum
141.8	33.79	5.81	140	132	154

Adding the same number 2 to each observation, then the summary statistics becomes

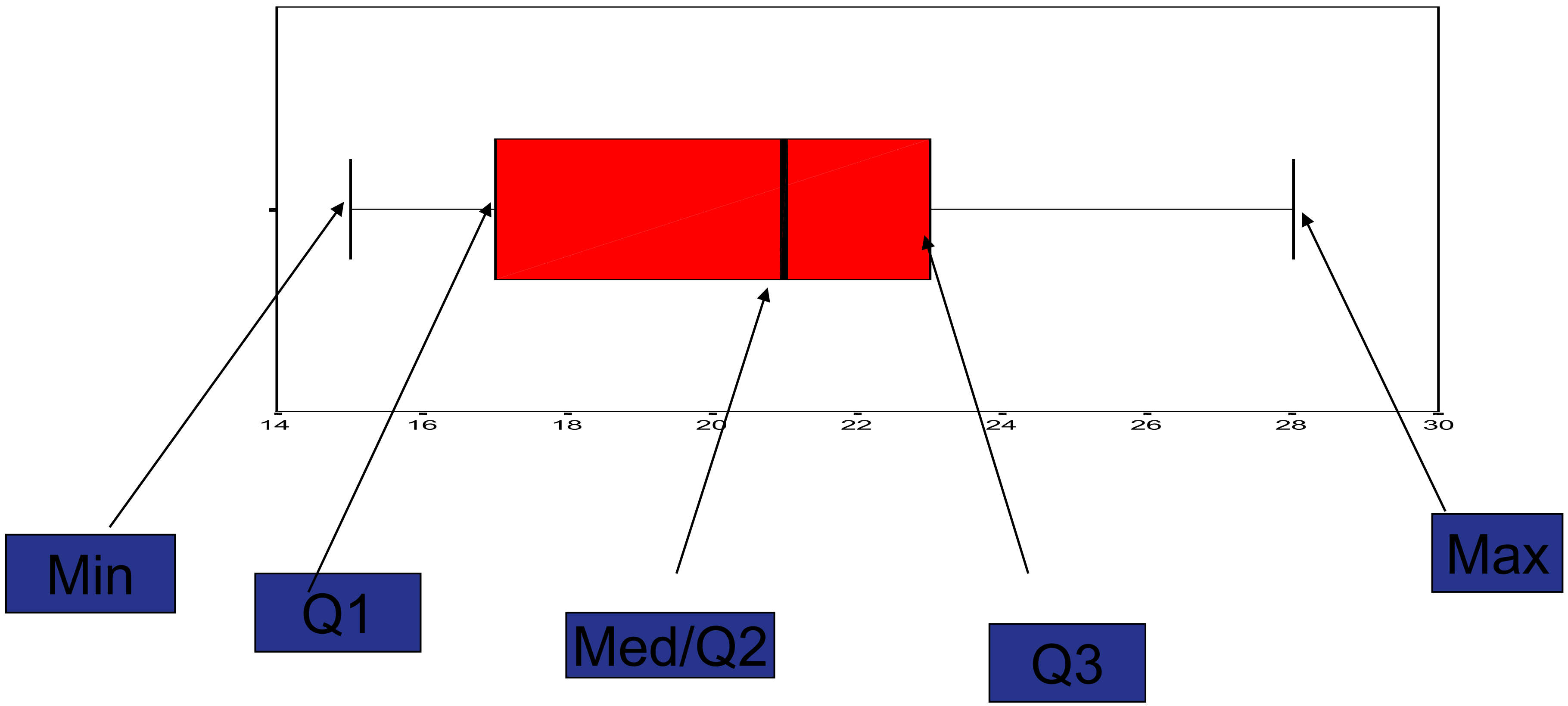
Sample Mean	Sample Variance	Standard deviation	Median	Minimum	Maximum
143.8	33.79	5.81	142	134	156

Multiplying the same number 2 to each observation, then the summary statistics becomes

Sample Mean	Sample Variance	Standard deviation	Median	Minimum	Maximum
283.6	135.16	11.62	280	264	308

Aim 2 Box plot

Provides **5-number summary**



Five-number summary and boxplot

25	6	6.1
24	5	5.6
23	4	5.3
22	3	4.9
21	2	4.7
20	1	4.5
19	6	4.2
18	5	4.1
17	4	3.9
16	3	3.8
15	2	3.7
14	1	3.6
13		3.4
12	6	3.3
11	5	2.9
10	4	2.8
9	3	2.5
8	2	2.3
7	1	2.3
6	6	2.1
5	5	1.5
4	4	1.9
3	3	1.6
2	2	1.2
1	1	0.6

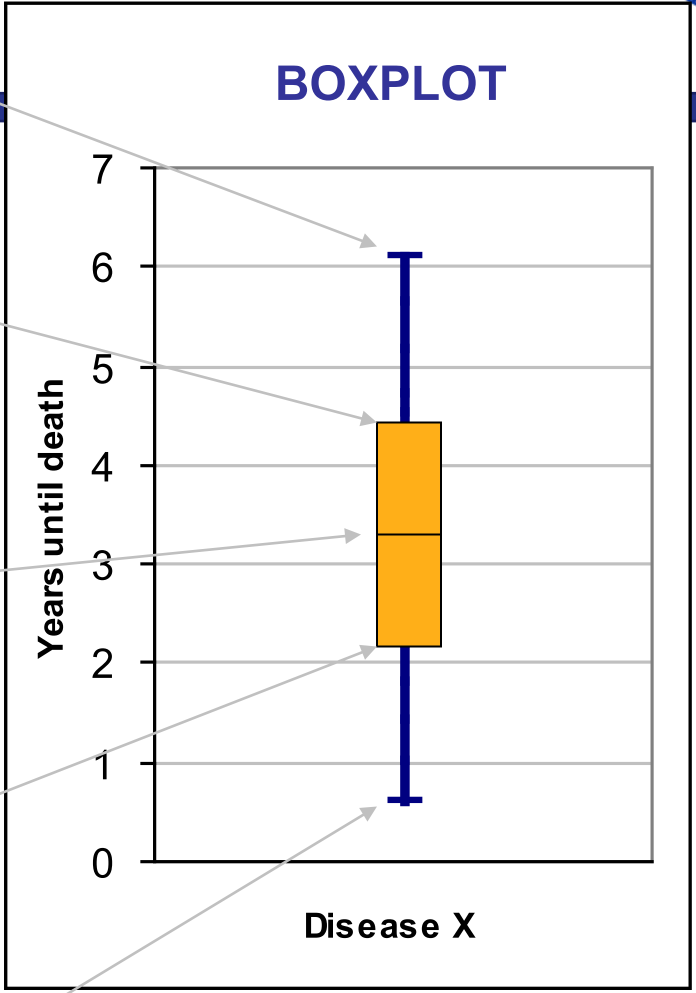
Largest = max = 6.1

Q_3 = third quartile = 4.35

M **= median = 3.4**

Q_1 = first quartile = 2.2

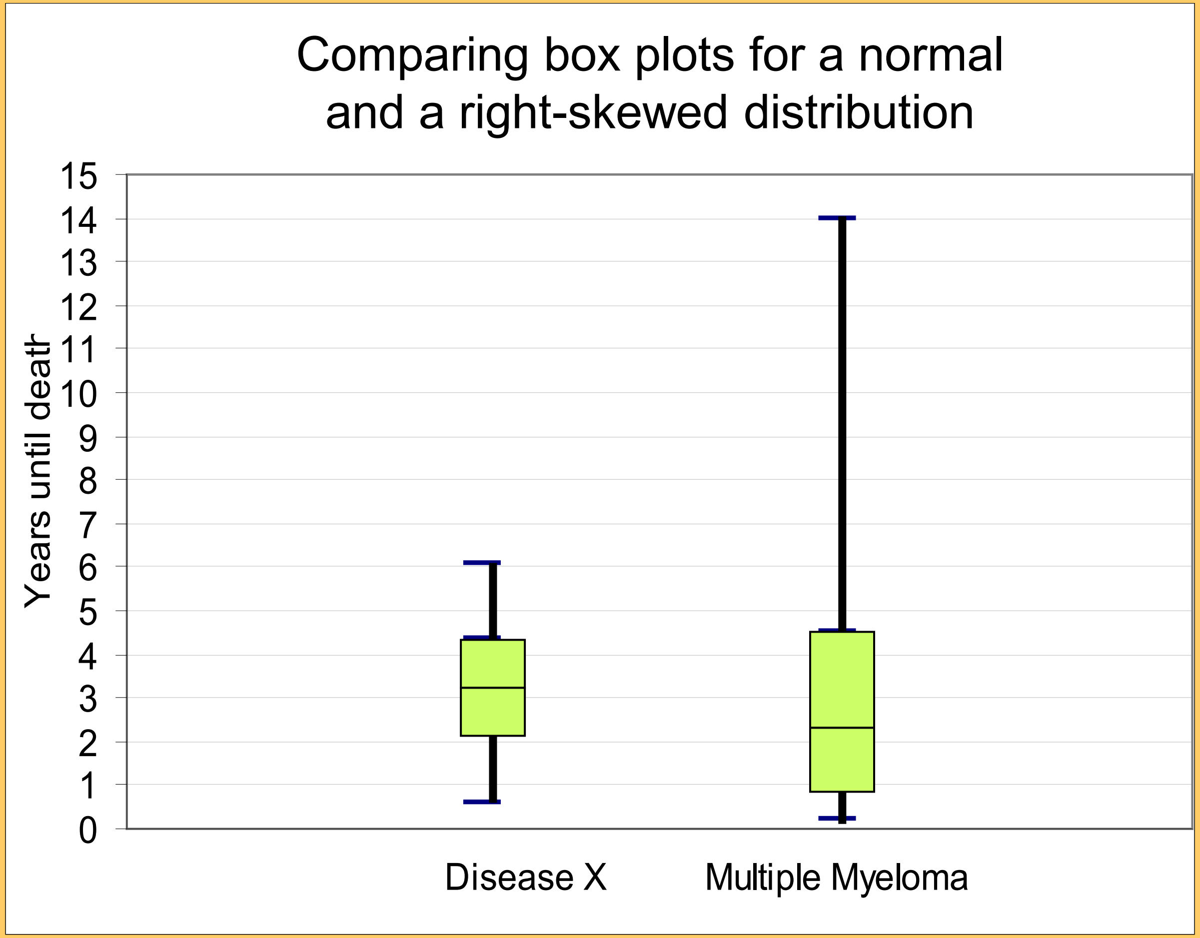
Smallest = min = 0.6

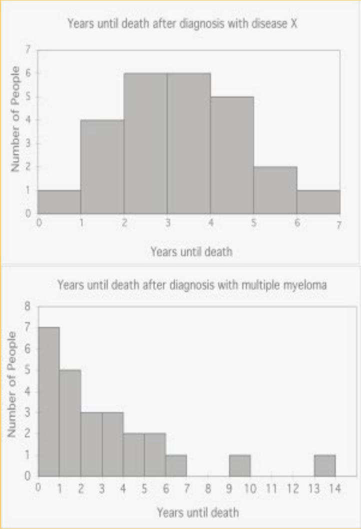


Five-number summary:

min Q_1 M Q_3 max

Boxplots for skewed data: Example 11





Boxplots remain true to the data and depict clearly symmetry or skew.

In Class Exercise 4.

If a distribution is skewed to the right, data taken from the distribution will tend to have a larger mean than median.

a) TRUE

b) FALSE

ANSWER

Skewed to the right

Some large values

Large values don’t affect for calculation of median

Large values will be used for mean (hence larger)

=>larger mean. TRUE

REVISION: Describing Distributions (Numerical data)

3 S's

Shape

Symmetric

Skewed to the right; Skewed to the left

(S)Center

Mean

Median – which to use?

Spread

Standard deviation; Range; IQR – which to use?

Aim 3 Data Screening and Outliers

In practice, data sets often contain errors.

They can occur for a variety of reasons, Eg:

- ✓Recording error in the field or laboratory;
- ✓Transcription error at the data entry stage;
- ✓Gross measurement error;
- ✓Inclusion of inappropriate experimental material;
- ✓Misinterpretation of recording instructions;
- ✓Change to definition of variables;
- ✓Missing value codes treated as data.

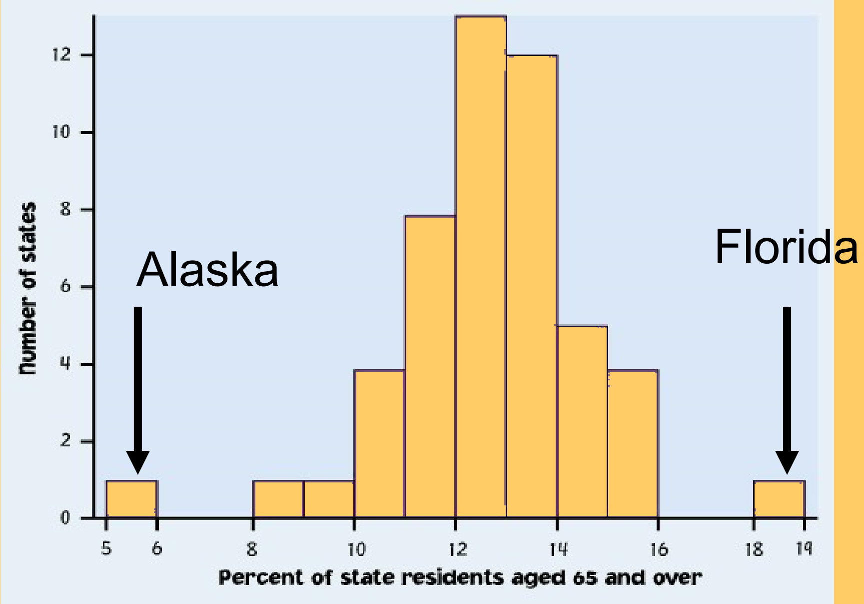
Outliers

An important kind of deviation is an **outlier.** Outliers are observations that lie outside the overall pattern of a distribution.

Always look for outliers and try to explain them.

The overall pattern is fairly symmetrical except for 2 states that clearly do not belong to the main group. Alaska and Florida have unusual representation of the elderly in their population.

A large gap in the distribution is typically a sign of an outlier.



Mild or suspected outliers

- One way to raise the flag for a suspected outlier is to compare the distance from the suspicious data point to the nearest quartile (Q_1 or Q_3). We then compare this distance to the **interquartile range** (distance between Q_1 and Q_3).

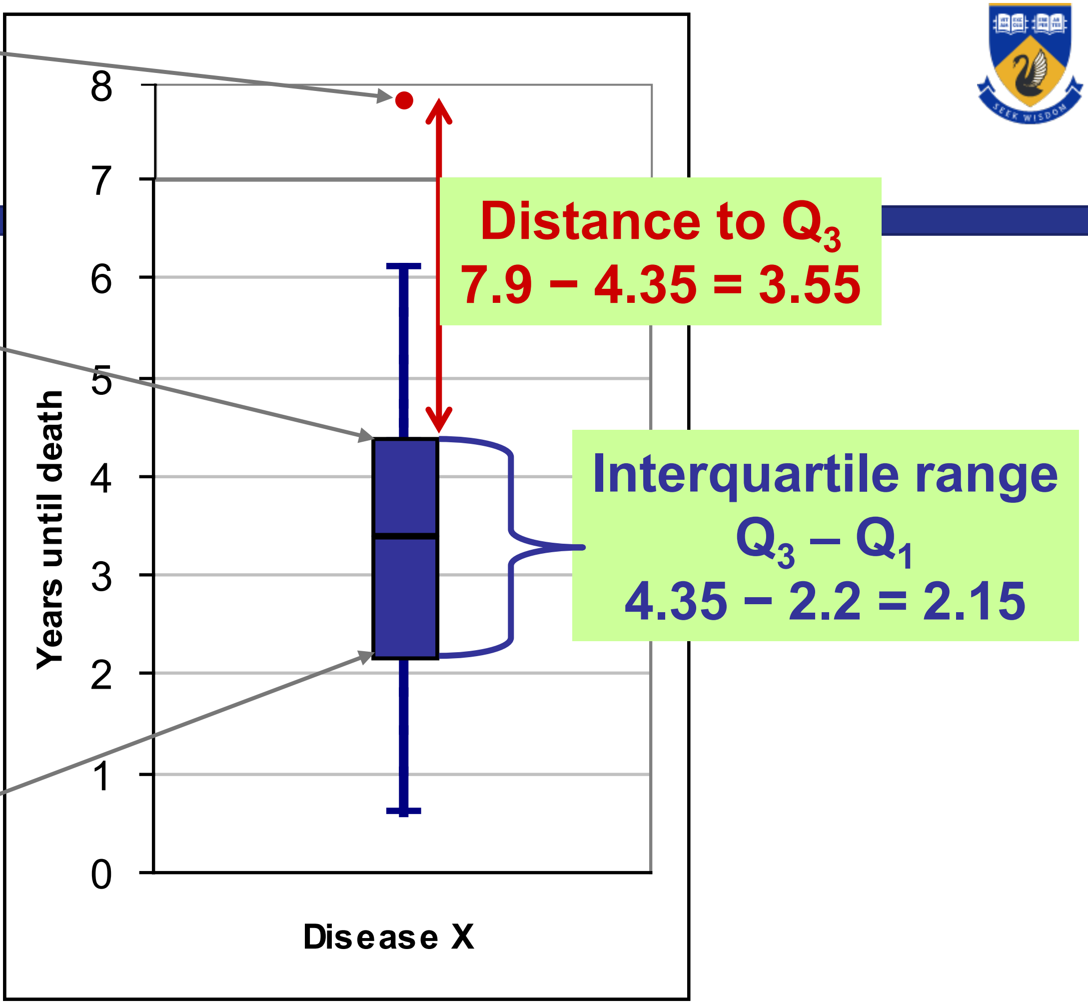
- We call an observation a **suspected outlier** if it falls more than 1.5 times the size of the interquartile range (IQR) above the first quartile or below the third quartile. This is called the "**1.5 * IQR rule for outliers.**"

THE UNIVERSITY OF WESTERN AUSTRALIA

25	6	7.9
24	5	6.1
23	4	5.3
22	3	4.9
21	2	4.7
20	1	4.5
19	6	4.2
18	5	4.1
17	4	3.9
16	3	3.8
15	2	3.7
14	1	3.6
13		3.4
12	6	3.3
11	5	2.9
10	4	2.8
9	3	2.5
8	2	2.3
7	1	2.3
6	6	2.1
5	5	1.5
4	4	1.9
3	3	1.6
2	2	1.2
1	1	0.6

$Q_3 = 4.35$

$Q_1 = 2.2$



Individual #25 has a value of 7.9 years, which is 3.55 years above the third quartile.

This is more than 1.5 * IQR=3.225 years. Thus, individual #25 is a mild/suspected outlier.

More on outliers

THE UNIVERSITY OF WESTERN AUSTRALIA

- If an outlier is confirmed as an error, and only then, should it be corrected or removed;
- Outliers are not necessarily errors:

Eg. death rate for young men in 1917 (WW1).

- Errors are not necessarily outliers:
  - ✓ recording the age 23 as 32 is unlikely to produce an outlier;
  - ✓ a research assistant who guesses a value to replace a failed measurement may not produce an outlier.

The BIG picture: Examining Relationships

Most statistical studies involve more than one variable.

Questions:

- What cases does the data describe?
- What variables are present and how are they measured?
- Are all of the variables quantitative?
- Do some of the variables explain or even cause changes in other variables?

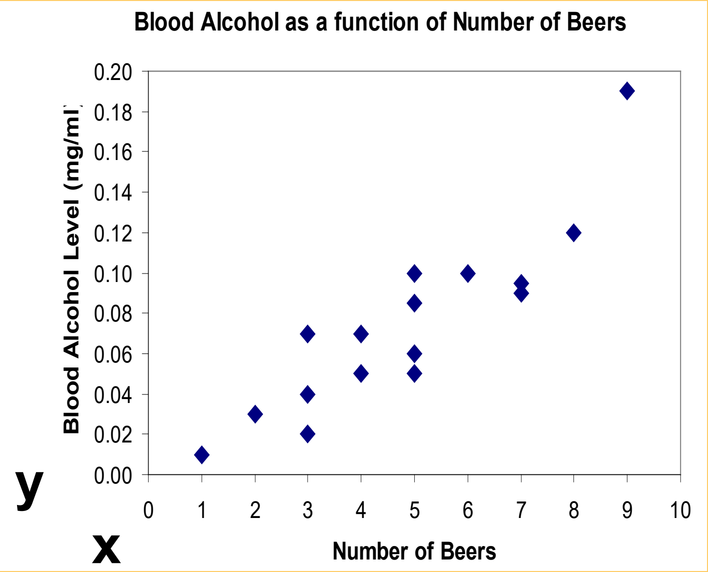
Aim 4 Explanatory and response variables

A **response variable** measures or records an outcome of a study.

An **explanatory variable** explains changes in the response variable.

Typically, the *explanatory or independent variable* is plotted on the x axis, and the *response or dependent variable* is plotted on the y axis.

Response (dependent) variable:
blood alcohol content



Explanatory (independent) variable:
number of beers

Relationships involving numerical variables

- (Y, response) Numerical variable and (X, explanatory) Categorical variable
  - Side by side box-plots
  - Alternatively, histograms

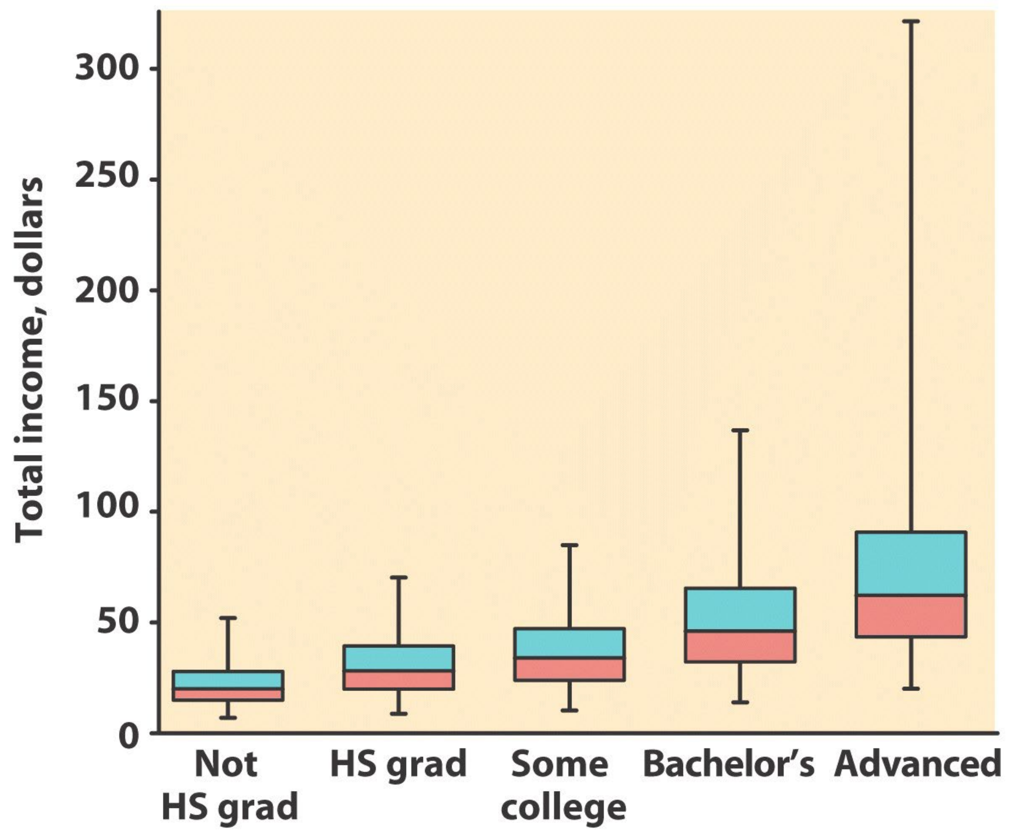
- (Y, response) Numerical variable and (X, explanatory) Numerical variable
  - Scatterplot

Aim 5 Relationship Between a Continuous Response Variable (Y) and a Categorical Explanatory (X) Variable: side-by-side boxplots

When the explanatory variable is categorical, you cannot make a scatterplot, but you can compare the different categories side by side on the same graph (boxplots, or mean +/– standard deviation).

Example 12 Comparison of income (quantitative response variable) for different education levels (categorical ordinal explanatory with five categories).

But be careful in your interpretation: This is NOT a positive association, because education is not quantitative.



Example 13: Forensic Science

A crime is committed. A few hairs are found at the scene. The hair is analysed in a laboratory and a particular hair conditioner additive, distearoylethyl hydroxyethylmonium (DH), is found.

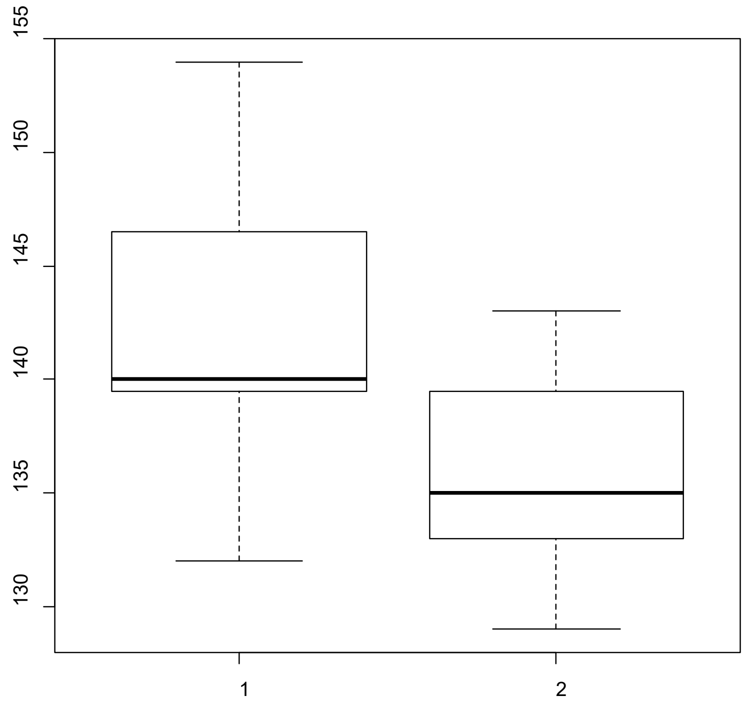
The viscosity of DH (in cP) is considered an important clue to the type of conditioner used in the hair.

Measurements of DH taken from the hair at the crime scene are given below.

146	154	141	140	136	132	147
140	147	139	140	140		

Viscosity measurements for DH in samples of the suspect's hair were also taken. These were as follows.

135	139	132	134	142	135	130
138	134	143	129			



1 "The viscosity of DH at the crime scene"

2 " The viscosity of DH of the suspect's hair"

Exploring data to find possible relationships

Different plots for different combinations of types of variables

Two example plots on the golf ball data (produced in lecture). The data columns are

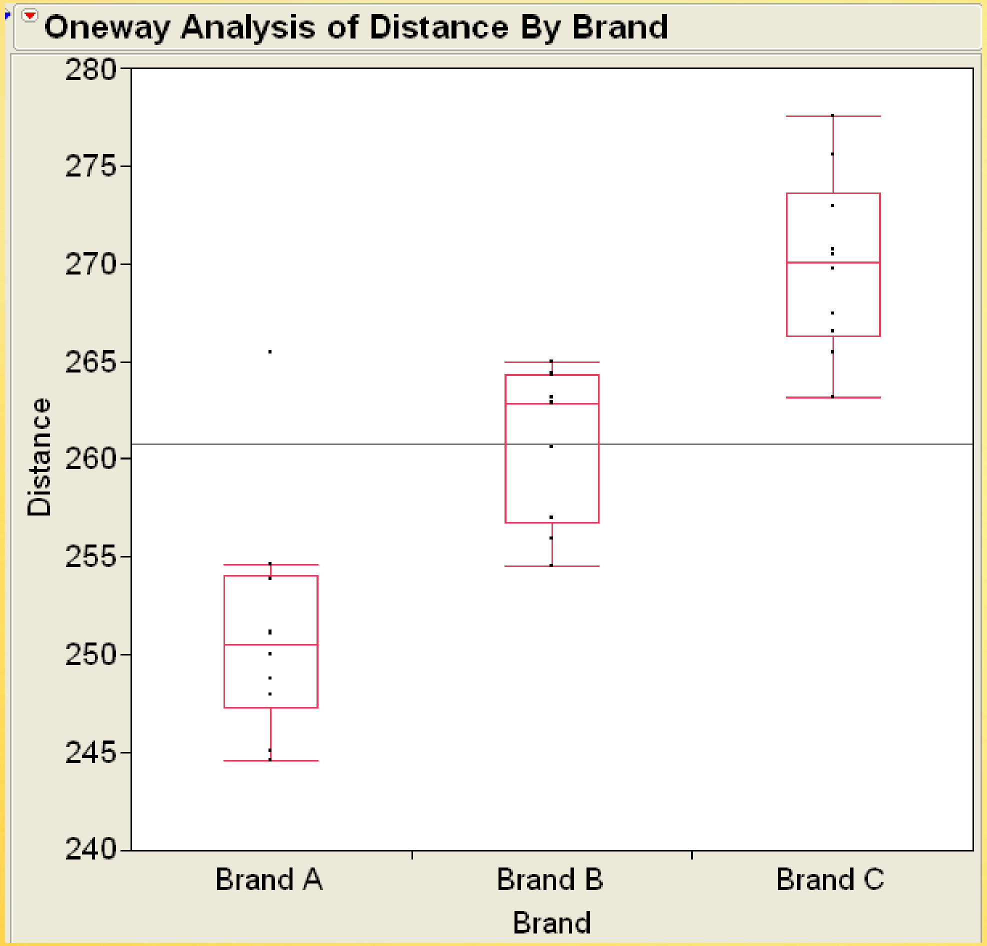
- Brand
- Distance (of flight of golf ball when hit by a robotic club)
- Durability measure

Example 14
Golf ball data

	Brand	Distance	Durability
1	Brand A	251.2	310
2	Brand B	263.2	261
3	Brand C	269.7	233
4	Brand A	245.1	235
5	Brand B	262.9	219
6	Brand C	263.2	289
7	Brand A	248.0	279
8	Brand B	265.0	263
9	Brand C	277.5	301
10	Brand A	251.1	306
11	Brand B	254.5	247
12	Brand C	267.4	264
13	Brand A	265.5	237
14	Brand B	264.3	288
15	Brand C	270.5	273
16	Brand A	250.0	284
17	Brand B	257.0	197
18	Brand C	265.5	208
19	Brand A	253.9	259
20	Brand B	262.8	207
21	Brand C	270.7	245
22	Brand A	244.6	273
23	Brand B	264.4	221
24	Brand C	272.9	271
25	Brand A	254.6	219
26	Brand B	260.6	244
27	Brand C	275.6	298
28	Brand A	248.8	301
29	Brand B	255.9	228
30	Brand C	266.5	276

In-Class Exercise 5.
What are the data types for Brand, Distance and Durability?

Distance (numerical response) **explained by** ***Brand*** (categorical explanatory)



In Class Exercise 6.
Start with 3S by comparing:

- Shape?
- (S) Center
- Spread?