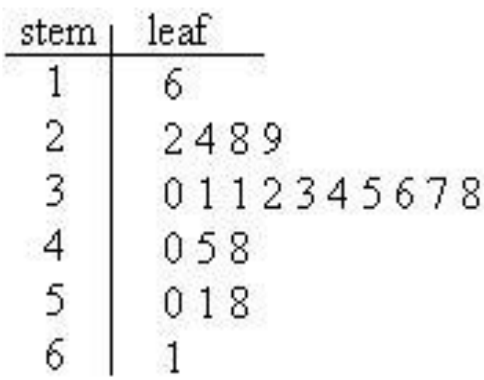
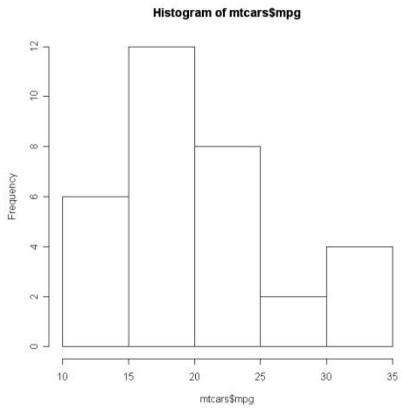
Lecture 2

Quantitative variables

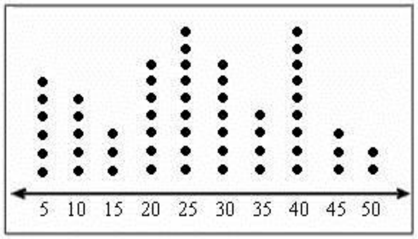
There are three main graphical methods for describing, summarizing, and detecting patterns in quantitative data:



- Stemplot (stem-and-leaf plot)



- Histogram



- Dot plot

Stemplots

stem	leaf
1	6
2	2 4 8 9
3	0 1 1 2 3 4 5 6 7 8
4	0 5 8
5	0 1 8
6	1

A *stemplot* gives a quick picture of the shape of a distribution while including the actual numerical values in the graph.

To make a stemplot:

1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf**, the final digit.

2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

StatCrunch -> Graphics -> Stem and Leaf

Example: Literacy of men and women: table below shows the percent of men and women at least 15 years old who were literate in 2002 in the major Islamic nations:

Country	Female percent	Male percent
Algeria	60	78
Bangladesh	31	50
Egypt	46	68
Iran	71	85
Jordan	86	96
Kazakhstan	99	100
Lebanon	82	95
Libya	71	92
Malaysia	85	92
Morocco	38	68
Saudi Arabia	70	84
Syria	63	89
Tajikistan	99	100
Tunisia	63	83
Turkey	78	94
Uzbekistan	99	100
Yemen	29	70

2 | 9
3 | 1 8
4 | 6
5 |
6 | 0 3 3
7 | 0 1 1 8
8 | 2 5 6
9 | 9 9 9

clusters

Variable: Female percent

Decimal point is 1 digit(s) to the right of the colon.

(means 2 | 9. = 29)

2 digits would be 290

3 . . . 2900

decimal point at the colon would give you 2.9

```
2 : 9
3 : 18
4 : 6
5 :
6 : 033
7 : 0118
8 : 256
9 : 999
```

Variable: Male percent

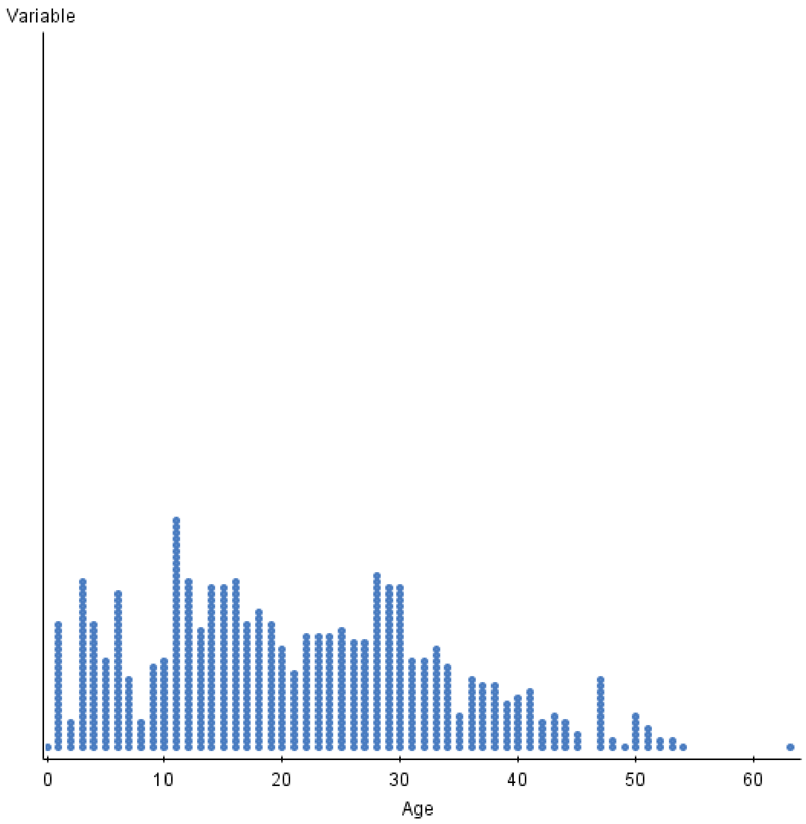
Decimal point is 1 digit(s) to the right of the colon.

```
 5 : 0
 6 : 88
 7 : 08
 8 : 3459
 9 : 22456
10 : 000
```

Dotplots

A **dotplot** is a simple display. It places a dot along an axis for each case in the data. It is very like a stemplot but with dots instead of digits for all the leaves. Dotplots show basic facts about distribution. They are quite useful for small data sets.

Here is a dotplot of ages for a group of people:



StatCrunch -> Graphics -> Dotplot

Examining a distribution:

• In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

• Overall pattern of a distribution can be described by its **shape**, **centre**, and **spread**.

• An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

• Some other things to look for in describing shape are:

- Does the distribution have one or several major peaks, usually called **modes**? A distribution with one major peak is called **unimodal**.

- Is it approximately symmetric or skewed in one direction?

Example: Describe the shapes of the distributions summarized by the following stemplots.

```
Stem-and-leaf of stab22 marks N = 42

6 | 7
7 | 44
7 | 77888999
8 | 00011233444
8 | 555556666778
9 | 000001
9 | 7
10| 0
```

symmetric
unimodal

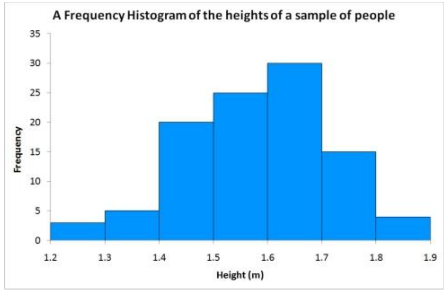
```
Stem-and-leaf of C1 N= 50

0 | 00011112223333444
0 | 555556666788999
1 | 0011444
1 | 5669
2 | 03
2 |
3 | 1
3 |
4 | 2
```

skewed to the right

possible outliers?

Histograms



• A histogram breaks the range of values of a variable into intervals and displays only the count or percent of the observations that fall into each interval.

• We can choose a convenient number of intervals.

• Histograms do not display the actual values observed (only counts in each interval).

<u>Example</u>: Here is some data on the number of days lost due to illness of a group of employees:

47, 1, 55, 30, 1, 3, 7, 14, 7, 66, 34, 6, 10, 5, 12, 5, 3, 9, 18, 45, 5, 8, 44, 42, 46, 6, 4, 24, 24, 34, 11, 2, 3, 13, 5, 5, 3, 4, 4, 1

The main steps in constructing a histogram

1. Determine the *range* of the data (largest and smallest values)
In our example:

2. Decide on the number of intervals (or classes), and the width of each class (usually equal).
3. Count the number of observations in each class. These counts are called class frequencies.
4. Draw the histogram.

47, 1, 55, 30, 1, 3, 7, 14, 7, 66, 34, 6, 10, 5, 12, 5, 3, 9, 18, 45, 5, 8, 44, 42, 46, 6, 4, 24, 24, 34, 11, 2, 3, 13, 5, 5, 3, 4, 4, 1

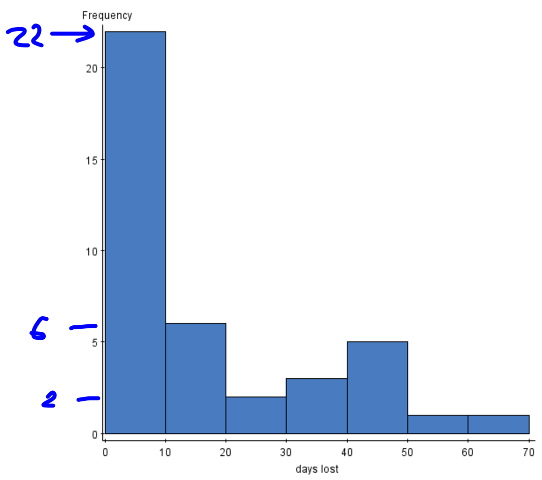
Class	# of employees (frequency)	Cumulative frequency	Relative frequency
0-9	22	22	0.55
10-19	6	28	0.15
20-29	2	30	0.05
30-39	3	33	0.075
40-49	5	38	0.125
50-59	1	39	0.025
60-69	1	40	0.025
Total	$\boxed{40}$		1.000

$= \frac{22}{40}$

Class	# of employees (frequency)	Cumulative frequency	Relative frequency
0-9	22	22	0.55
10-19	6	28	0.15
20-29	2	30	0.05
30-39	3	33	0.075
40-49	5	38	0.125
50-59	1	39	0.025
60-69	1	40	0.025
Total	40		1.000

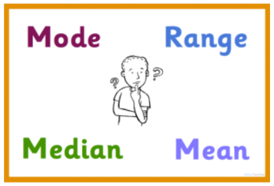
• A table with the first two columns above is called **frequency table** or **frequency distribution**.

• A table with the first column and the third column is called **cumulative frequency distribution**.



StatCrunch -> Graphics -> Histogram

Describing distributions with numbers:



A large number of numerical methods are available for describing quantitative data sets. Most of these methods measure one of two data characteristics:

- The **central tendency** or **location** of the set of observations – that is the tendency of the data to cluster, or center, about certain numerical values.

- The **variability** of the set of observation – that is the spread of the data.

Measuring Center

Two common measures of center are

- **mean** ("average value")

- **median** ("middle value")

Mean:

To find the **mean** $\bar{x}$ of a set of observations, add their values and divide by the number of observations. If the *n* observations are $x_1, x_2, \ldots, x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

sigma

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

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

Example:

Fuel economy (miles per gallon) for 2004 model vehicles

Two-Seater Cars			Minicompact Cars		
Model	City	Highway	Model	City	Highway
Acura NSX	17	24	Aston Martin Vanquish	12	19
Audi TT Roadster	20	28	Audi TT Coupe	21	29
BMW Z4 Roadster	20	28	BMW 325CI	19	27
Cadillac XLR	17	25	BMW 330CI	19	28
Chevrolet Corvette	18	25	BMW M3	16	23
Dodge Viper	12	20	Jaguar XK8	18	26
Ferrari 360 Modena	11	16	Jaguar XKR	16	23
Ferrari Maranello	10	16	Lexus SC 430	18	23
Ford Thunderbird	17	23	Mini Cooper	25	32
Honda Insight	60	66	Mitsubishi Eclipse	23	31
Lamborghini Gallardo	9	15	Mitsubishi Spyder	20	29
Lamborghini Murcielago	9	13	Porsche Cabriolet	18	26
Lotus Esprit	15	22	Porsche Turbo 911	14	22
Maserati Spyder	12	17			
Mazda Miata	22	28			
Mercedes-Benz SL500	16	23			
Mercedes-Benz SL600	13	19			
Nissan 350Z	20	26			
Porsche Boxster	20	29			
Porsche Carrera 911	15	23			
Toyota MR2	26	32			

hybrid comes from different population

$\rightarrow n = 21$

Let's find the mean highway mileage for two-seaters:

$$\bar{x} = \frac{24+28+28+\ldots+32}{21} = \frac{518}{21} = 24.7$$

Remove 66:

$$\bar{x} = \frac{24+28+\ldots+32}{20} = 22.6$$

Weakness of the mean: it is sensitive to the influence of a few extreme observations (outliers, skewed distribution). We say that it is NOT a **resistant measure** of center.

Median:

The **median** is the midpoint of the distribution, the number such that half the observations are smaller than it and the other half are larger.

To find the median of a distribution:

1. Arrange the observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median is the center observation in the ordered list.
3. If the number of observations n is even, the median is the average of the two center observations in the ordered list.

$\frac{n+1}{2}$

Example: The annual salaries (in thousands of $) of a random sample of five employees of a company are: 40, 30, 25, 200, 28

Arranging the values in increasing order: 25 28 30 40 200

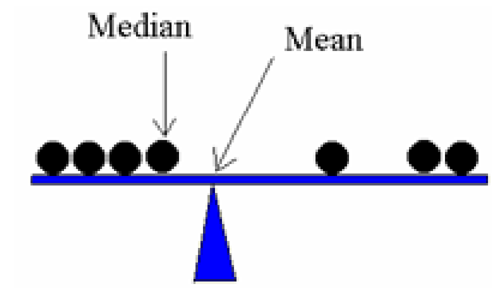
$\frac{n+1}{2} = 3$

median = 30

Excluding 200 median = $\frac{28+30}{2} = 29$

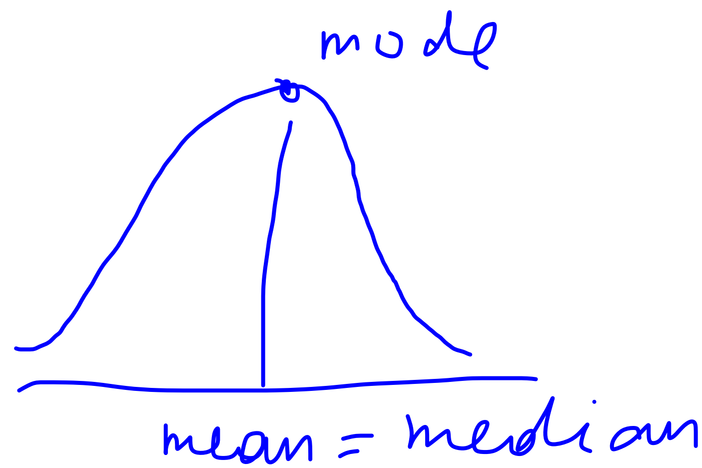
Note: the median is more resistant than the mean.

Mean versus median:



• The median and mean are the most common measures of the center of a distribution.

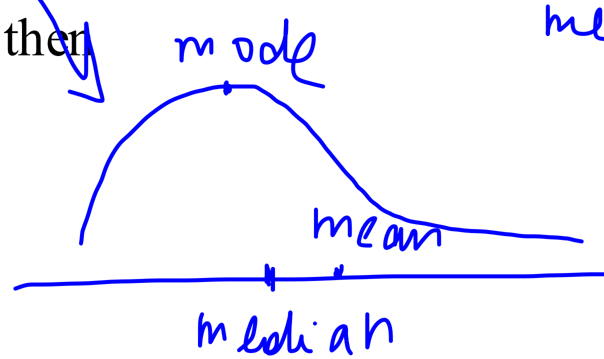
• If the distribution is exactly symmetric, the mean and median are exactly the same.



• Median is less influenced by extreme values.

• If the distribution is skewed to the right, then

mode < median < mean



• If the distribution is skewed to the left, then

mean < median < mode

Trimmed mean:

• Trimmed mean is a measure of the center that is more resistant than the mean but uses more of the available information than the median.
• To compute the 10% trimmed mean, discard the highest 10% and the lowest 10% of the observations and compute the mean of the remaining 80%. Similarly, we can compute 5%, 20%, etc. trimmed mean.
• Trimming eliminates the effect of a small number of outliers.

Example: Compute the 10% trimmed mean of the data given below.

```
20 40 22 22 21 21 20 10 20 20 20 13 18 50 20 18 15 8 22 25
```

Solution:

- Arrange the values in increasing order: $\bar{X}$

```
8 10 13 15 18 18 20 20 20 20 20 20 21 21 22 22 22 25 40 50
```

- There are 20 observations and 10% of 20 = 2. Hence, discard the first 2 and the last 2 observations in the ordered data and compute the mean of the remaining 16 values.

```
Mean = 19.812
```

Measuring Spread

There are two main measures of spread that we will discuss: range and standard deviation.

The **range** (max-min) is a measure of spread but it is very sensitive to the influence of extreme values.

The measure of spread that is used most often is the **standard deviation.**

The **variance** s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \dots, x_n$ is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

or, in more compact notation,

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

The **standard deviation s** is given by

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

It can be shown that, $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$ ← skip this

This formula is usually quicker.

The idea behind the variance and the standard deviation as measures of spread is as follows:

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

• The deviations $x_i - \bar{x}$ display the spread of the values x_i about their mean. Some of these deviations will be positive and some negative because the observations fall on each side of the mean.

• The sum of the deviations of the observations from their mean will always be zero.

$\boxed{(x_1 - \bar{x})} + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0$

$= (x_1 + x_2 + \ldots x_n) - \bar{x} - \bar{x} - \ldots - \bar{x} = (x_1 + \ldots + x_n) - n \cdot \bar{x} = 0$

• Squaring the deviations makes them all positive, so that observations far from the mean in either direction have large positive squared deviations.

• The variance is the average of the squared deviations.

• The variance, s^2, and the standard deviation, s, will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

Example: Find the standard deviation of the following data set:

4, 8, 2, 9, 7 $\quad n = 5$ – sample size

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

$$s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{4+8+2+9+7}{5} = \frac{30}{5} = 6$$

$$s^2 = \frac{(4-6)^2 + (8-6)^2 + (2-6)^2 + (9-6)^2 + (7-6)^2}{5-1}$$

$$= 8.5$$

$$s = \sqrt{8.5} = \underline{2.92}$$

Properties of standard deviation:

• *s* measures the spread about the mean and should be used only when the mean is chosen as the measure of center.

• $s = 0$ only when there is no spread. This happens only when all observations have the same value. Otherwise, $s > 0$.

5 5 5 5 5 $\Rightarrow \bar{x} = 5$

• *s*, like the mean, is not resistant to extreme values. A few outliers can make *s* very large.

Ballpark approximation for *s*:

The ballpark approximation for the standard deviation *s* is the Range divided by 4 (divide by 3 if there are less than 10 observations, divide by 5 if there are more than 100 observations).

<u>Example</u>: for the data set

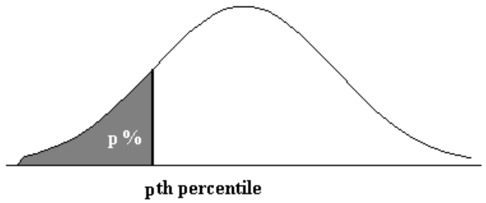
range $= 9 - 2 = 7$

4, 8, 2, 9, 7

$5 < 10$

$s \approx \frac{7}{3} = 2.33$

Percentiles



• The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.

• We can describe the spread by giving several percentiles.

• The $\boldsymbol{p^{th}}$ **percentile** of a distribution is the value such that *p* percent of the observations are smaller or equal to it.

<u>Example</u>: the median is the 50th percentile.

• If a data set contains *n* observations, then the *p*th percentile is the $\boxed{(n+1) \times \frac{p}{100}^{th}}$ value in the ordered data set.

Example: Find the 20th percentile of the data represented by the following stem-and-leaf plot.

```
Stem-and-leaf of Rural N = 29

2 | 1
3 | 3589
4 | 122333456788
5 | 112467
6 | 7
7 | 04
8 | 48
9 |
10| 8
```

$n = 29$

$p = 20$

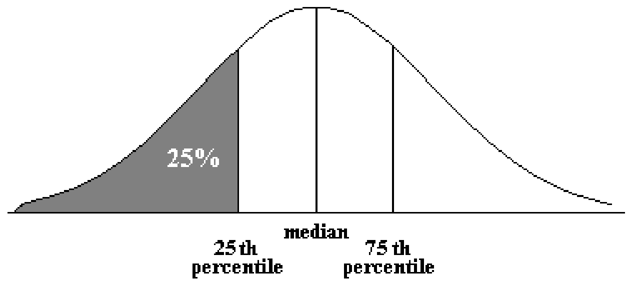
$$(n+1)\frac{p}{100} = (29+1)\frac{20}{100}$$

$$= 30 \cdot \frac{20}{100} = 6$$

20^{th} percentile $= 6^{th}$ obs'n

$= 41$

Quartiles



• The 25th percentile is called the **first quartile** (Q_1).

• The first quartile (Q_1) is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

• The 75th percentile is called the **third quartile** (Q_3).

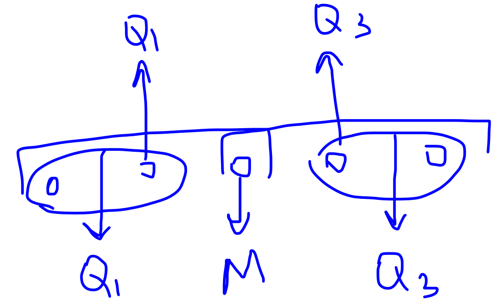
• The third quartile (Q_3) is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

<u>Note</u>: The median is the second quartile (Q_2).

M

Example: The highway mileages of 20 cars, arranged in increasing order are:

13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32.



The median is … $\dfrac{23+23}{2} = 23$

The first quartile Q_1 is … $\dfrac{17+19}{2} = 18$

The third quartile Q_3 is… $\dfrac{26+28}{2} = 27$

The Five-Number Summary

The **five-number summary** of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \quad Q_1 \quad \text{M} \quad Q_3 \quad \text{Maximum}$$

These five numbers give a reasonably complete description of both the center and the spread of the distribution.

Example: The highway mileages of 20 cars, arranged in increasing order are:

13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32

Give the five-number summary.

StatCrunch -> Stat -> Summary Stats

Summary statistics:

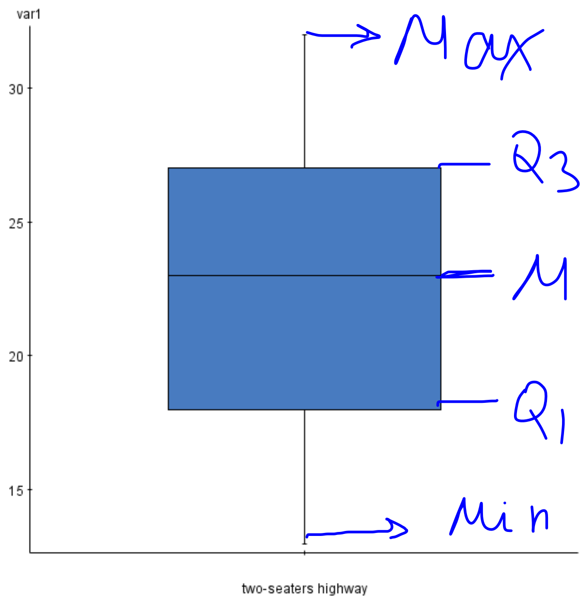
Column	n	Mean	Variance	Std. Dev.	Median	Range	Min	Max	Q1	Q3
var1	20	22.6	27.936842	5.2855315	23	19	13	32	18	27

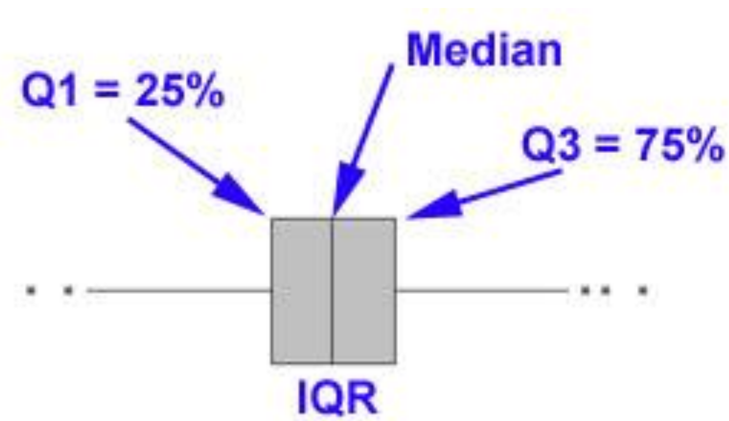
Answer:

Min	Q1	M	Q3	Max
13	18	23	27	32

Boxplots:

A **boxplot** is a graph of the five-number summary:





- A central box spans the quartiles Q_1 and Q_3.

- A line in the box marks the median M.

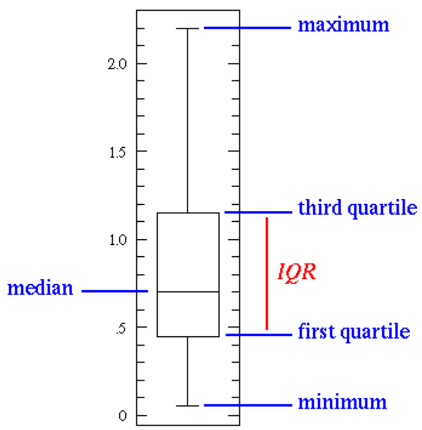
- Lines extend from the box out to the smallest and largest observations.

StatCrunch -> Graphics -> Boxplot

IQR

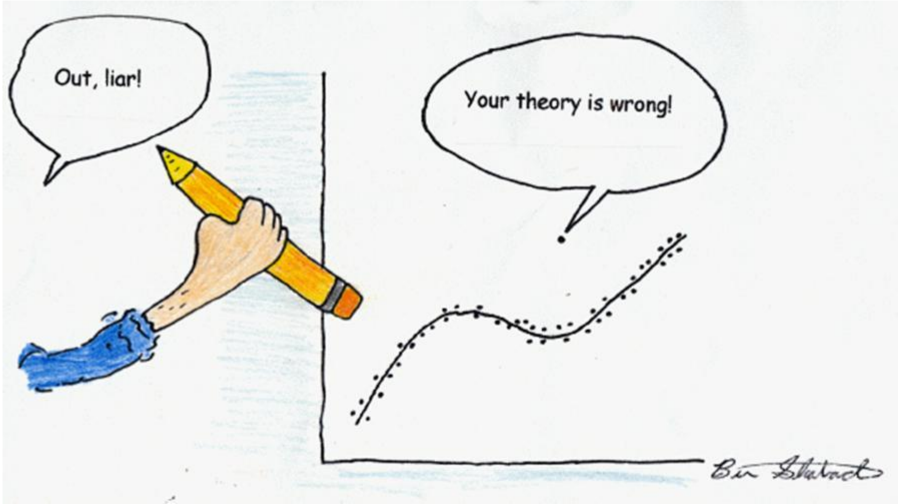
• The **range** (max-min) is a measure of spread but it is very sensitive to the influence of extreme values.

• The distance between the first and third quartiles is called the **interquartile range** (IQR) i.e. $IQR = Q_3 - Q_1$.



• The IQR is another measure of spread that is less sensitive to the influence of extreme values, like

Outliers

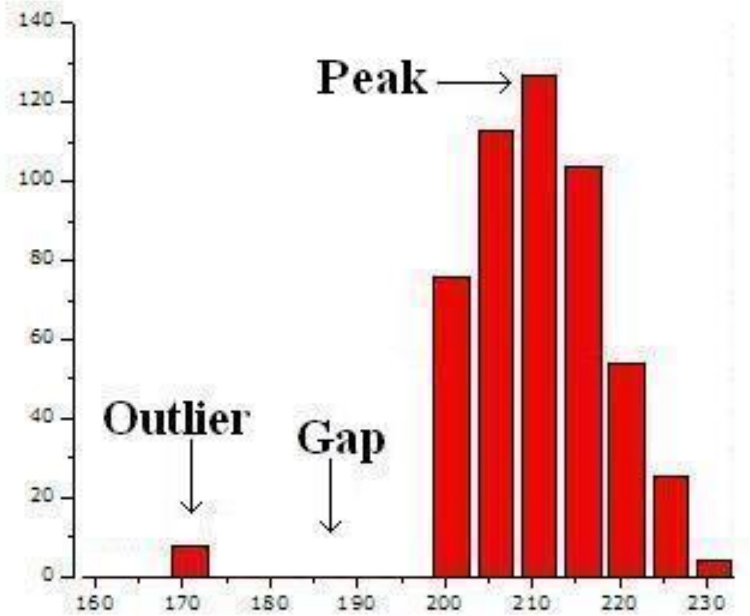


An **outlier** is an observation that is usually large or small relative to the other values in a data set. Outliers are typically attributable to one of the following causes:

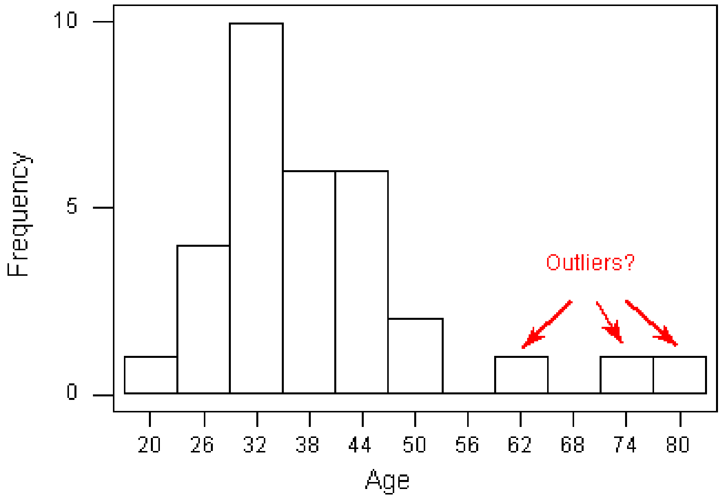
1. The observation is observed, recorded, or entered incorrectly.

2. The observation comes from a different population.

3. The observation is correct but represents a rare event.



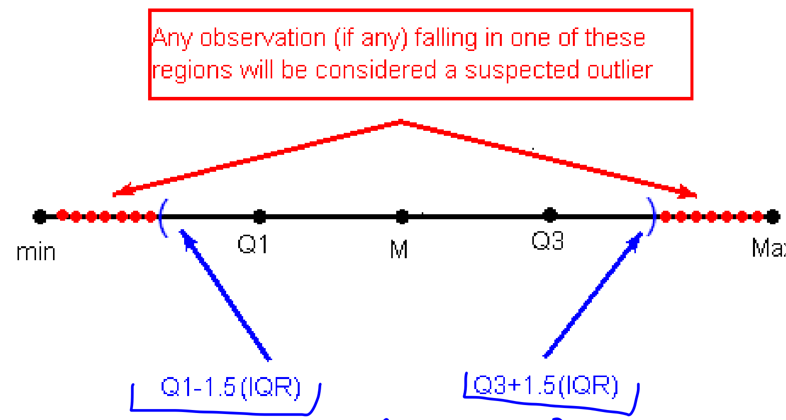
The *1.5×IQR* Criterion for outliers



Call an observation a suspected outlier if it falls

- more than *1.5×IQR* above the 3rd quartile or

- more than *1.5×IQR* below the 1st quartile.



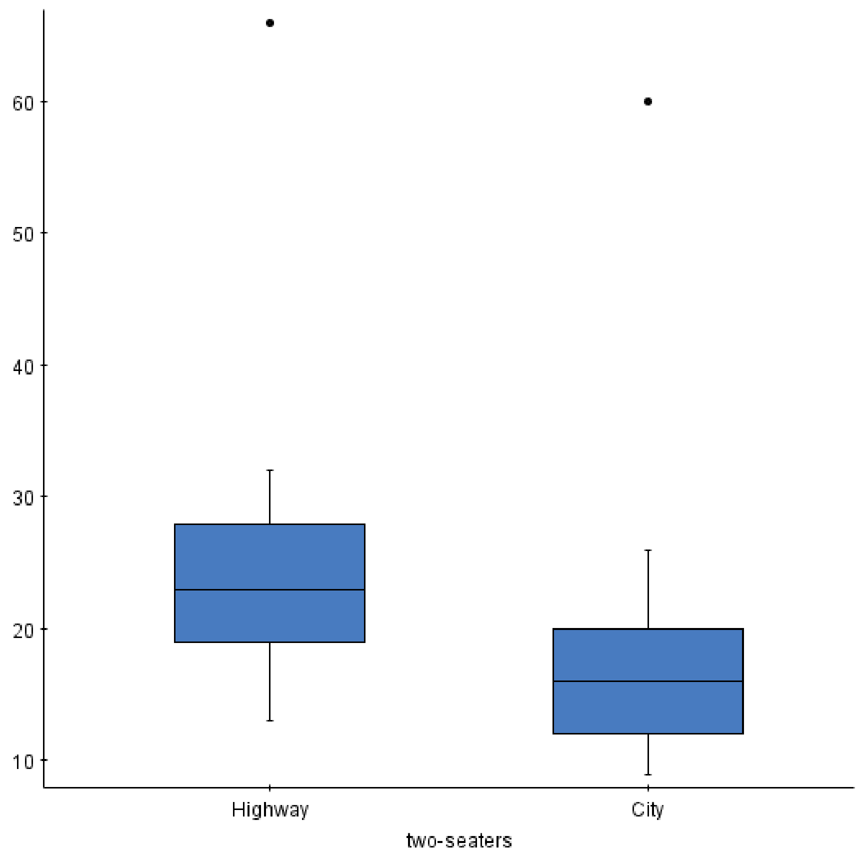
$IQR = Q_3 - Q_1$

$(Q_1 - 1.5\,IQR,\ Q_3 + 1.5 \cdot IQR)$

Example: Consider the data given in the previous example (mileage data with an extra observation of 66).

Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Median	Range	Min	Max	Q1	Q3	IQR
var1	21	24.666666	116.23333	10.781157	23	53	13	66	19	28	9



$IQR = Q_3 - Q_1 = 28 - 19$
$= 9$

$Q_1 - 1.5 \cdot IQR$
$= 19 - 1.5 \cdot 9 = 19 - 13.5 = 5.5$

Note: $\min = 13 > 5.5$

$Q_3 + 1.5 \cdot IQR$
$= 28 + 13.5 = 41.5$

$\boxed{66 > 41.5}$

So 66 is an outlier
↳ Honda Insight (hybrid)

<u>Example</u>: Find five-number summary from the given stemplot.

```
Stem-and-leaf of stab22 marks N = 42

6 | 7
7 | 44
7 | 77888999
8 | 0001123344
8 | 55555666778
9 | 000001
9 | 7
10| 0
```

Min	Q_1	M	Q_3	Max
67	79	84	87	100

$$M = \frac{21^{st} + 22^{nd}}{2} = \frac{84+84}{2} = 84$$

$Q_1 = 11^{th} = 79$ or $Q_1 = (42+1)\frac{25}{100}^{th} = 10.75^{th} \approx 11^{th}$

$(n+1)\frac{p}{100}$

$Q_3 = 11^{th}$ from the end or $(21+11)^{th} = 32^{nd} = 87$

or $Q_3 = (42+1)\frac{75}{100} = 32.25^{th} \approx 32^{nd}$ obs'n