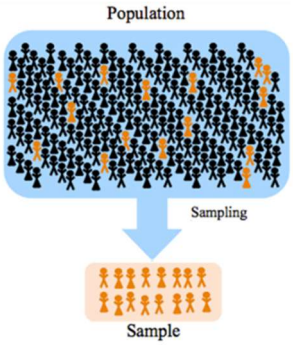
2 – Descriptive Statistics: Numerical Measures

Statistics and Parameters

In statistics we are often interested in a variable X that takes a definite value for each *individual* (or *unit*) in a *population*.

e.g. $X = \text{Age of a BCIT student}$

Quite often, we cannot measure X for all individuals in the population. We can only determine X for a *sample* (i.e., a subset) of the population.



Numbers that summarize X for the whole population are called *parameters*.

Numbers that summarize X for a sample are called *statistics*.

In this course you will be working with the following statistics and parameters.

		Statistic Symbol	**Parameter Symbol**
Measures of Proportion (Categorical)	• proportion		
Measures of Centre	• Mean • Median		
Measures of Position	• Quartile • Percentile		
Measures of Variation	• Standard Deviation • Range		

2.1 – Measure of Proportion (Non-Numerical Data)

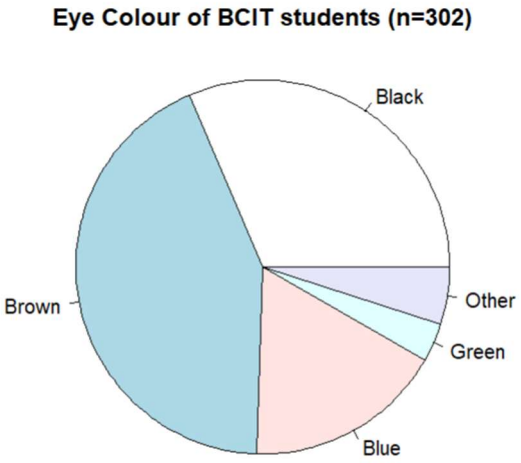
For categorical (i.e., non-numerical) variables, we can only *count* the number of individuals in a certain category and determine the *proportion* of individuals in that category.

sample proportion: $\hat{p} = \frac{X}{n}$

population proportion: $p = \frac{X}{N}$

Example Suppose a sample of BCIT students gives the following counts for the categorical variable $X = \text{Eye Colour}$. Determine the sample proportion of Brown-eyed students.

Eye Colour	**Count (# Students)**
Black	95
Brown	130
Blue	52
Green	10
Other	15



2.2 - Measures of Centre

A *measure of centre* for a variable X is a single numerical value which indicates the "typical" or "central" value of X.

Arithmetic Mean

The most common measure of centre is the *arithmetic mean*.

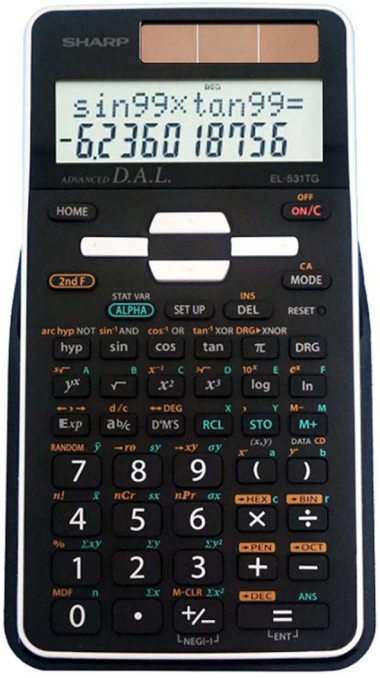
Sample mean: $\bar{X} = \frac{\Sigma X}{n}$ (statistic) n is the sample size

Population mean: $\mu = \frac{\Sigma X}{N}$ (parameter) N is the population size

Example Calculate the mean of the first 10 values for eruption times for Old Faithful:

3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950 4.350

$\bar{X} =$



You can use your calculator's statistical functions to compute the mean:

In R, the command to find $\bar{X}$ is:

```
mean(faithful$eruptions)
```

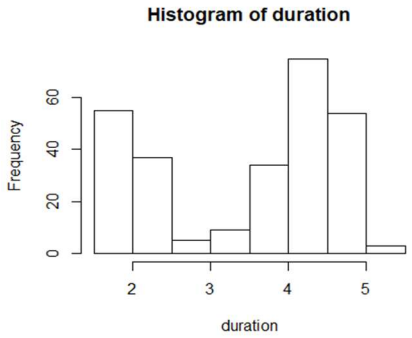
Mean of Grouped Data

Some technical documents present data not in its *raw* form but rather only as *grouped* data. In that case, you cannot calculate $\bar{X} = \frac{\Sigma X}{n}$. Instead, you must use

$$\bar{X} = \frac{\Sigma\,[f_i X_i]}{\Sigma f_i}$$

- i is the index of the groups ($i = 1, 2, 3, \ldots$)
- $f_i =$ the number of X values in group i
- $X_i =$ the *class mark* (midpoint) of group i

Example Suppose we have just the following frequency distribution for Old Faithful eruptions. The class marks and frequencies are shown.



X_i	1.75	2.25	2.75	3.25	3.75	4.25	4.75	5.25
f_i	51	41	5	7	30	73	61	4

$\bar{X} =$

In R, we could use the commands:

```
> Xi <- c(1.75, 2.25, 2.75, 3.25, 3.75, 4.25,4.75, 5.25)
> fi <- c(51, 41, 5, 7, 30, 73, 61, 4)
> X.bar <- sum(Xi * fi) / sum(fi)
```

Median

The *median* of a set of X values is the middle value when the values are sorted from least to greatest. The median is denoted by $\tilde{X}$ or Q_2.

- for an *odd* number of data values, the median is located at the exact middle

- for an *even* number of data values, the median is the mean of the two middle values

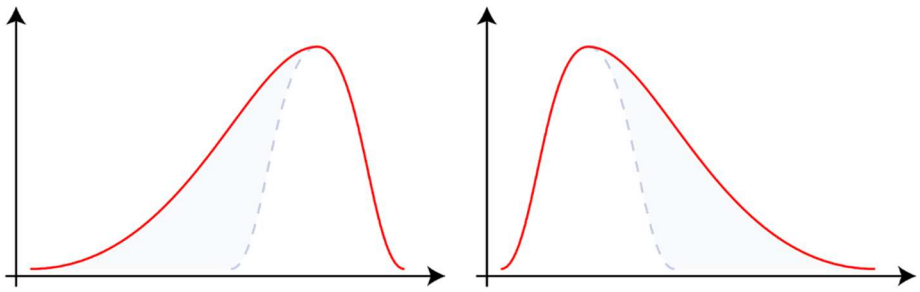
Example The first ten values of Old Faithful eruptions in sorted order is

1.800 1.950 2.283 2.883 3.333 3.600 3.600 4.350 4.533 4.700

So, the median of this sample is: $Q_2 =$

In R, the median is found using: `> median(faithful$eruptions)`

Note: If the distribution of X is *symmetric*, then its mean and median are equal. Otherwise, the mean and the median of a sample can be quite different.

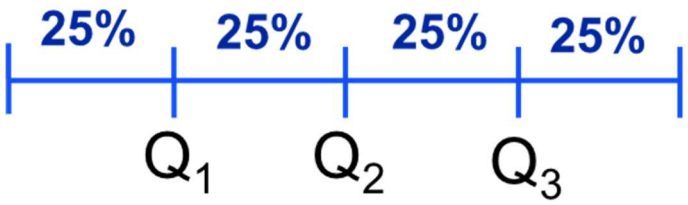


2.3 - Measures of Position

The median value of X tells us what value is halfway through the dataset. More generally, a *measure of position* tells us what value of X stands at a certain place within the data set.

Quartiles, Percentiles, Quantiles

The *quartiles* of a data set are the values that divide it into four quarters (after sorting).



Example The variable $X =$ `faithful$eruptions` has $n = 272$ observations. 25% of 272 is 68. Therefore:

$Q_1 =$ the value that separates the lower 68 values and the upper 204 values

$Q_2 =$ the value that separates the lower 136 values and the upper 136 values

$$= \frac{(136\text{th value} + 137\text{th value})}{2} = 4.000$$

$Q_3 =$ the value that separates the lower 204 values and the upper 68 values

$$= \frac{(204\text{th value} + 205\text{th value})}{2} = 4.4585$$

In R, we use the commands:

- `Q1 <- quantile(faithful$eruptions, 0.25)`
- `Q2 <- quantile(faithful$eruptions, 0.5)`
- `Q3 <- quantile(faithful$eruptions, 0.75)`

The concept of *percentiles* generalizes the concept of quartiles. The kth *percentile* P_k is the value that separates the lower $k\%$ of a data set and the upper $(100 - k)\%$.

There are a few different formulas for determining P_k which give slightly different results. (The reason is that the formulas must *interpolate* between the values of X and this is done is different ways.) For our purposes, those differences are not important.

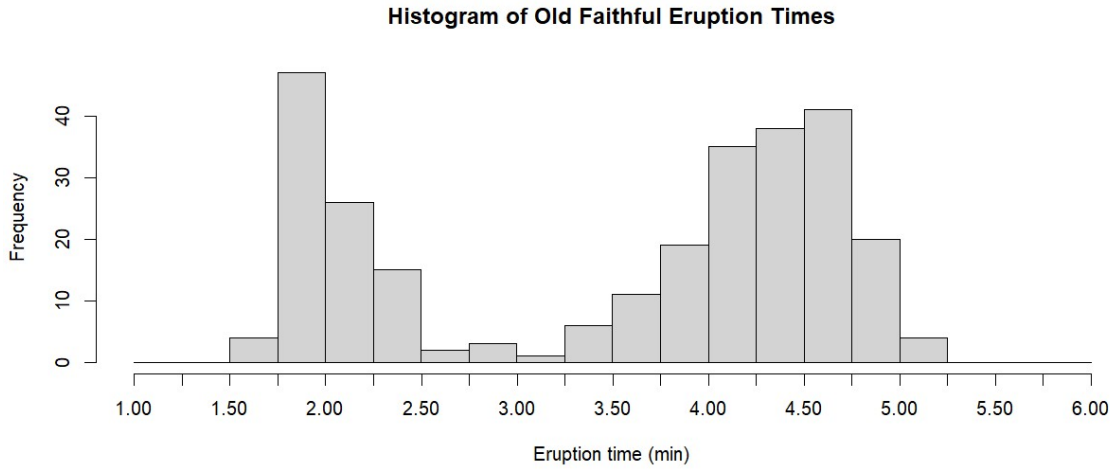
In R, the kth percentile is

```
Pk <- quantile( faithful$eruptions, k/100 )
```

Example The 33nd percentile of eruption times for Old Faithful is

```
> quantile(faithful$eruptions, 0.33)
   33%
2.417
```

This means 33% of eruption times are below 2.417 minutes (and 67% are above).



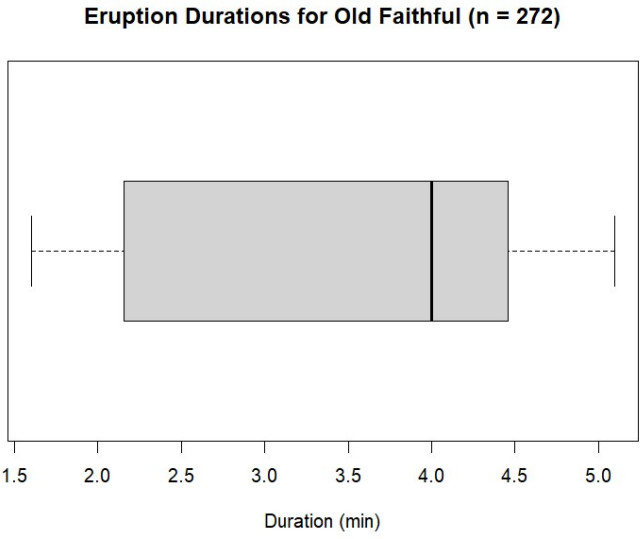
Box and Whisker Plot

If we use `quantile()` without specifying the fraction, R gives a summary of all quartiles:

```
> quantile(faithful$eruptions)
     0%      25%      50%      75%     100%
1.60000 2.16275 4.00000 4.45425 5.10000
```

These values are called the *five-number summary* of X. A *boxplot* (or *box-and-whisker* plot) is a visualization of the five-number summary.

Example A boxplot of eruption times for Old Faithful is shown below. The box contains the "middle 50%" of the values, since it covers the range from Q_1 to Q_3. The "whiskers" then extend out to the minimum and maximum values.



Note that the part of the box between Q_1 and Q_2 is quite large, which means that eruption times in that quarter of the data set are quite spread out.

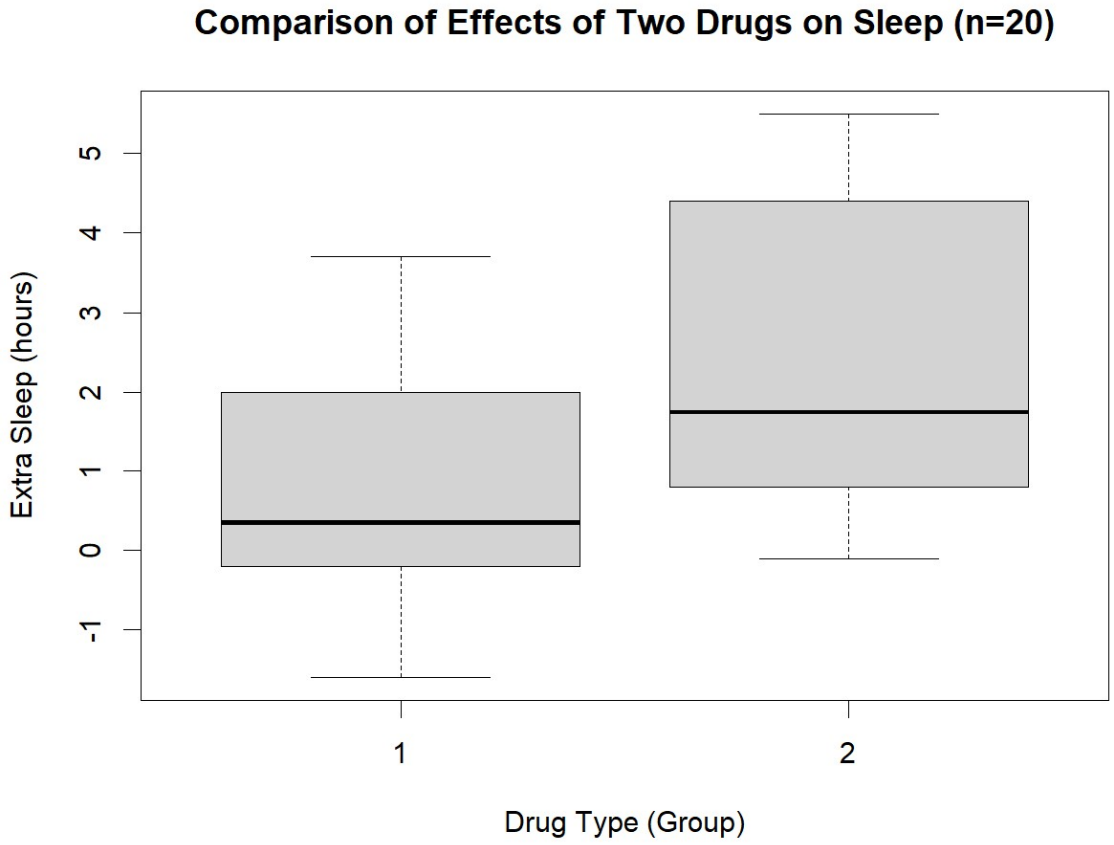
Although single boxplots by themselves are informative, probably the most common use of the boxplot is in comparing two or more sets of data.

Example Two types of drugs were tested on 10 students who had trouble sleeping. The built-in data-frame `sleep` provides results.

- variable `group` indicates which of the two drugs (listed here as just "1" and "2") the student received.
- variable `extra` gives the number of extra hours of sleep that each student had

We can get a visual representation of the effectiveness of the two types of drugs by creating side-by-side boxplots, grouping the students according to which drug they received.

```
> boxplot(extra~group, data=sleep)
```



What can we say about the two types of drugs based on this data?

Outliers

A difficult issue in statistical work is the question of what to do about outliers (suspiciously unusual numbers, either very large or very small compared to other numbers).

- Are the unusual values the results of errors?
- Should we delete outliers?

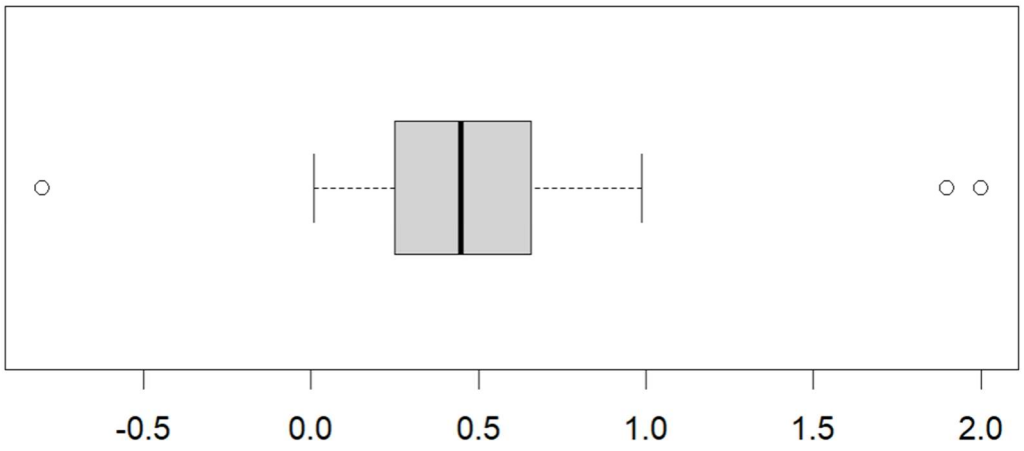
There is no absolute answer to these questions, since the human beings who gathered the data could have made a mistake they didn't know about ("unknown unknowns").

For the purpose of *identifying* outliers, statisticians use a rule based on the quartile values of a variable.

$$IQR = Q_3 - Q_1 \qquad \text{(}\textit{inter-quartile range}\text{)}$$

$$\text{Lower Fence} = \text{Q}_1 - 1.5 \times IQR$$

$$\text{Upper Fence} = \text{Q}_3 + 1.5 \times IQR$$

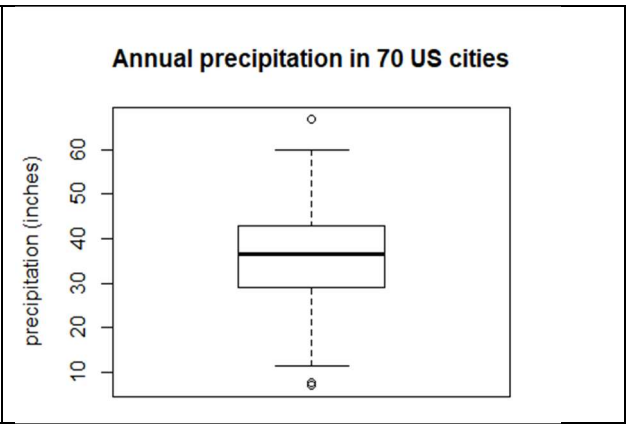


Any value that is below the lower fence or above the upper fence is an *outlier*.

On a boxplot, outlier values are shown using a separate symbol (a small circle or star). The whiskers then extend only to the min/max values that are between the two fences.

Example The dataframe *precip* gives annual rainfalls in 70 American cities.

```
> precip
          Mobile           Juneau          Phoenix
            67.0             54.7              7.0
     Little Rock      Los Angeles       Sacramento
            48.5             14.0             17.2
   San Francisco          Denver          Hartford
            20.7             13.0             43.4
      Wilmington      Washington      Jacksonville
            40.2             38.9             54.5
           Miami         Atlanta          Honolulu
            59.8             48.3             22.9
           Boise         Chicago           Peoria
            11.5             34.4             35.1
    Indianapolis      Des Moines          Wichita
            38.7             30.8             30.6
```



Note that the highest value (67 inches, for Mobile), is much higher than the second highest value (59.8 inches, for Miami). By representing the larger value as an outlier, we can clearly see how far it is from the rest of the data.

2.4 - Measures of Variation

We often need to measure how spread out the values of X are. For this we calculate *measures of spread* (or *variation*). There are several:

Range

$R =$ maximum value $-$ minimum value

Interquartile Range (IQR)

$IQR =$ range of "middle 50%" $= Q_3 - Q_1$

What advantage does the IQR have over the range?

What disadvantage does the IQR have over the range?

Example The interquartile range of Old Faithful eruptions is

$$IQR = Q_3 - Q_1 = 4.45 \text{ min} - 2.16 \text{ min} = 2.29 \text{ min}$$

Standard Deviation

The most important measure of spread is the *standard deviation*. It indicates the *typical* distance of X values away from the mean value.

Sample Standard Deviation $$s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$$ (statistic)

Population Standard Deviation $$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$ (parameter)

Example Calculate s from the following small sample of X values.

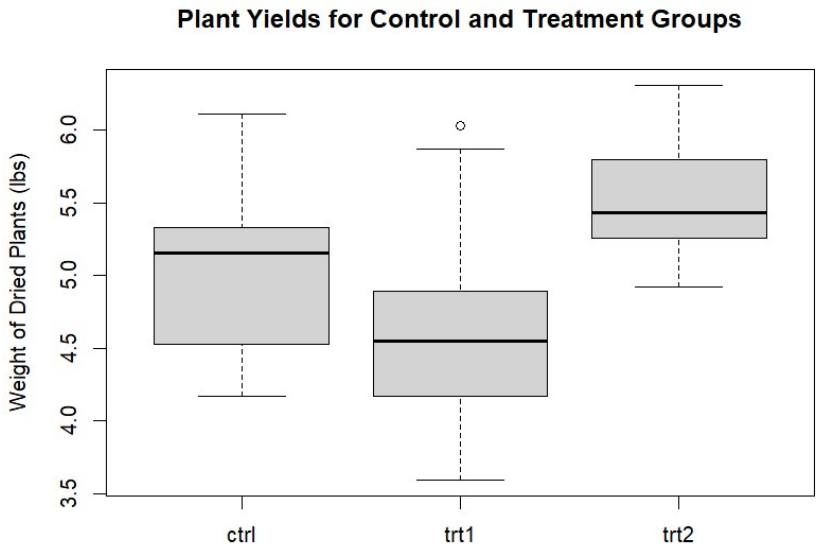
5.0 4.5 6.0 7.0 5.2

Note:

- Sample standard deviation s uses $n-1$ instead of n.
  The reason is that this makes s a better approximation of σ.
- Sample standard deviation s uses sample mean $\bar{X}$.
  Population standard deviation σ uses population mean μ.
- The units of s and σ are the same as the units of X.
- Standard deviation depends on *all* of the data values (unlike R and IQR).

Example (Plant Growth) The dataframe `PlantGrowth` gives results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions. We can compare the treatments with boxplots:

	weight	group
1	4	ctrl
2	6	ctrl
3	5	ctrl
4	6	ctrl
5	4	ctrl
6	5	ctrl
7	5	ctrl
8	5	ctrl
9	5	ctrl
10	5	ctrl
11	5	trt1
12	4	trt1
13	4	trt1



Which of the three datasets appears to have the largest standard deviation (most spread)?

We can compute the three standard deviations in R using:

```
# Plants in ctrl group
> ctrl.plants <- filter( PlantGrowth, group=="ctrl")
> sd( ctrl.plants$weight )
# Find standard deviation for all groups at once
> sd(data=PlantGrowth, weight~group)
      ctrl      trt1      trt2
0.5830914 0.7936757 0.4425733
```

(The units are _____.)

Which of the three groups had the most consistent weight results?

Coefficient of Variation

The standard deviation s (or σ) gives the *absolute* amount of variation. Sometimes, however, it is more appropriate to measure *relative* variation using the *coefficient of variation*:

$$CV = \frac{s}{\bar{X}} \times 100\%$$

$$CV = \frac{\sigma}{\mu} \times 100\%$$

Example Are students in MATH 3042 more spread out in terms of *height* or in terms of *age*? Suppose we collected data and found that:

For $X =$ height: $\bar{X} = 165$ cm and $s = 5.2$ cm.

For $X =$ age: $\bar{X} = 20.2$ yr and $s = 3.5$ yr.

The standard deviation for age is a smaller number ($3.5 < 5.2$) but it is greater *relative* to the typical student's age. Let's calculate to *coefficient of variation* for each variable:

$$CV(\text{height}) = \frac{s}{\bar{X}} \times 100\% = \frac{5.2 \text{ cm}}{165 \text{ cm}} \times 100\% = 3.15\%$$

$$CV(\text{age}) = \frac{s}{\bar{X}} \times 100\% = \frac{3.5 \text{ yr}}{20.2 \text{ yr}} \times 100\% = 17.33\%$$

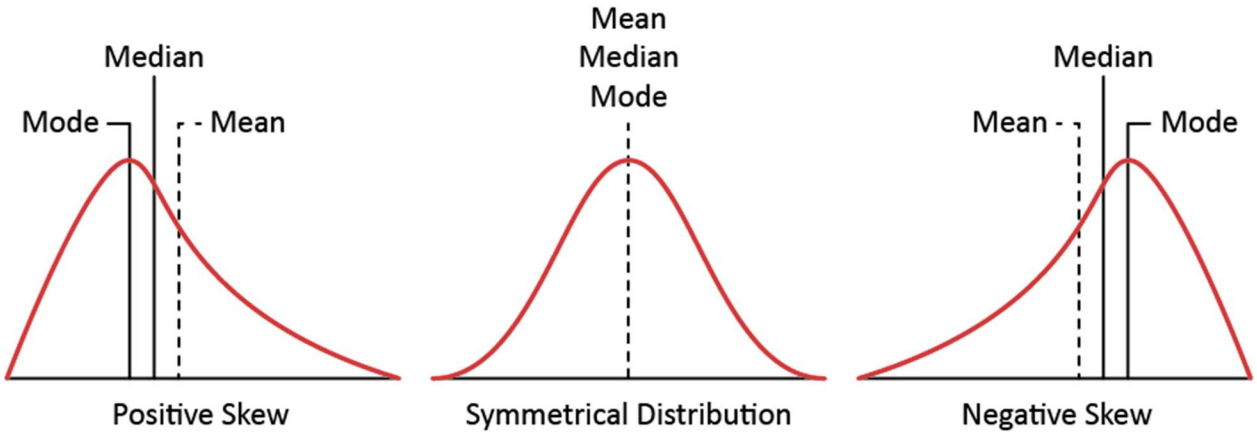
2.5 - Skewness (Sk)

Skewness, denoted Sk, is a numerical measure that does not fit into any of the previous categories (perhaps it is closest to being a measure of *position*). Skewness characterizes how *non-symmetrical* a variable is.

Pearson's Coefficient of Skewness

$$Sk = 3 \cdot \frac{\bar{X} - Q_2}{s}$$

$$Sk = 3 \cdot \frac{\mu - Q_2}{\sigma}$$

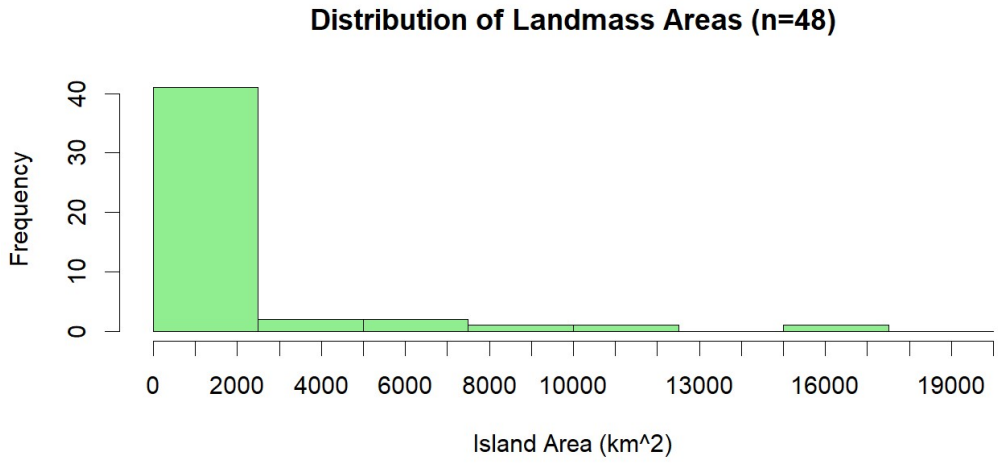


If X has a symmetric distribution, its mean and median are equal. $Sk \approx 0$

If X is skewed left, its mean is less than its median (due to the "left tail"). $Sk < -1$

If X is skewed right, its mean is greater than its median (due to the "right tail"). $Sk > +1$

Example The dataset `islands`, which lists the 48 largest land masses, is an example of a skewed dataset. There are a very small number of huge landmasses, such as North America, and many much smaller ones, such as Vancouver Island.



The skewness for area is:

$$Sk = 3 \cdot \frac{\bar{X} - Q_2}{s} = 1.078$$

```
> 3*(mean(islands)-median(islands))/sd(islands)
[1] 1.078324
```

Note: there are alternative formulas to measure skewness. The R function `skewness()`, which is part of the `moments` package, uses the formula

$$\text{skewness}(X) = \frac{\frac{1}{N}\sum_i (X_i - \mu)^3}{\left(\frac{1}{N}\sum_i (X_i - \mu)^2\right)^{\frac{3}{2}}}$$

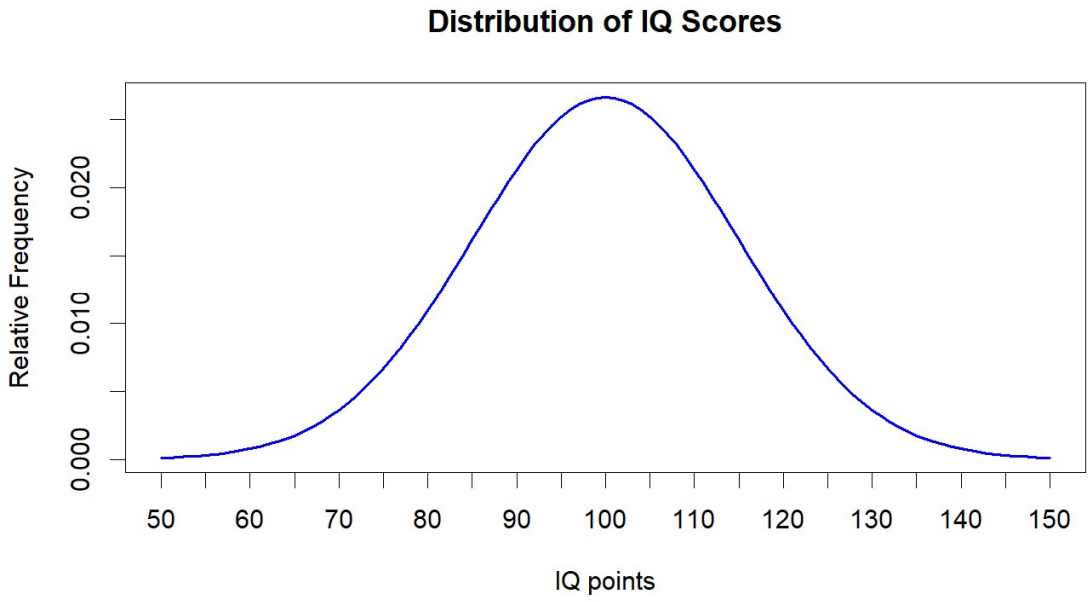
This formula does a better job than Pearson's formula (the first skewness formula) of capturing how skewed a dataset "looks", but it's laborious to compute. If you're using R, you should use the function `skewness()`, but if you are working with a calculator (on the MATH 3042 exam, for instance), then Pearson's formula is adequate.

2.6 - The Meaning of Standard Deviation

For variables that follow a *normal distribution*, the standard deviation has a very specific meaning and interpretation.

Example IQ scores follow a *normal* distribution with mean $\mu = 100$ and standard deviation σ. (Individual IQ scores tend to be about 15 units away from the central score of 100.)

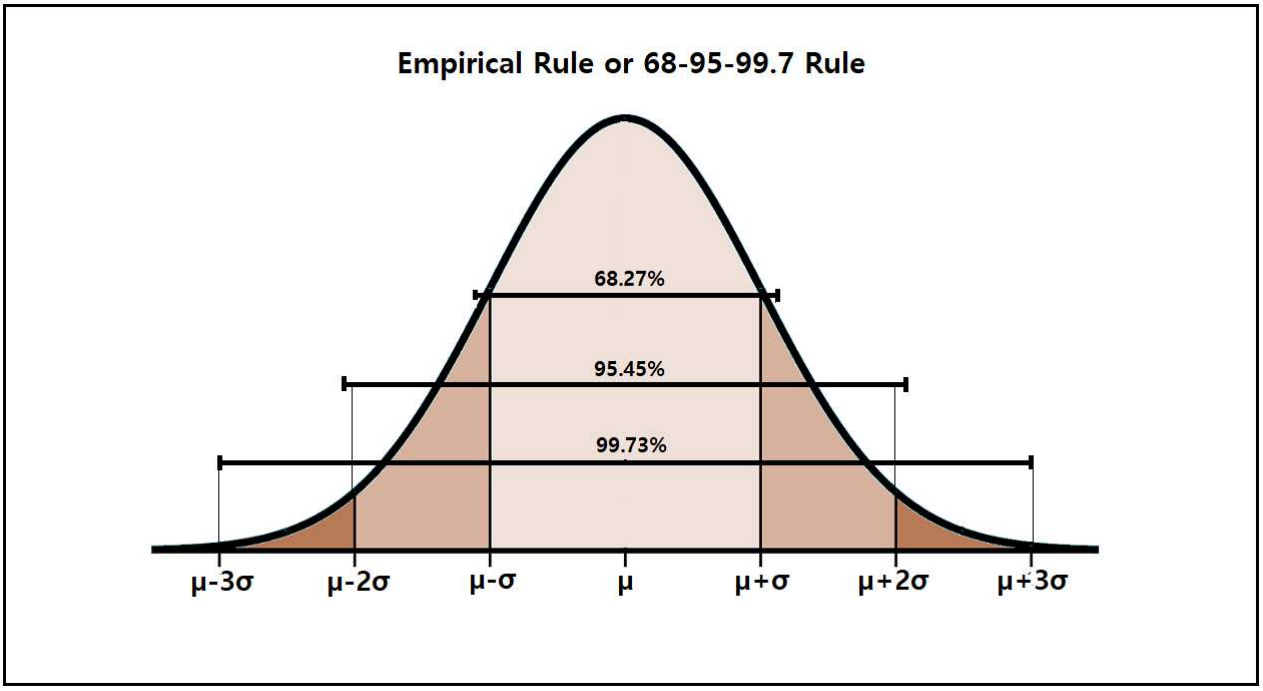
The shape of this distribution guarantees that 68% of individuals are "within one standard deviation of the mean", meaning: $85 \leq IQ \leq 115$



Empirical Rule

In general, if X follows a normal distribution, then the fraction of individuals with X between

$\mu - \sigma$	and	$\mu + \sigma$	is
$\mu - 2\sigma$	and	$\mu + 2\sigma$	is
$\mu - 3\sigma$	and	$\mu + 3\sigma$	is



Example What interval of IQ scores must contain 95% of all individuals?

Z Scores

The Z-score (or *standard score*) of an individual with respect to variable X is defined as:

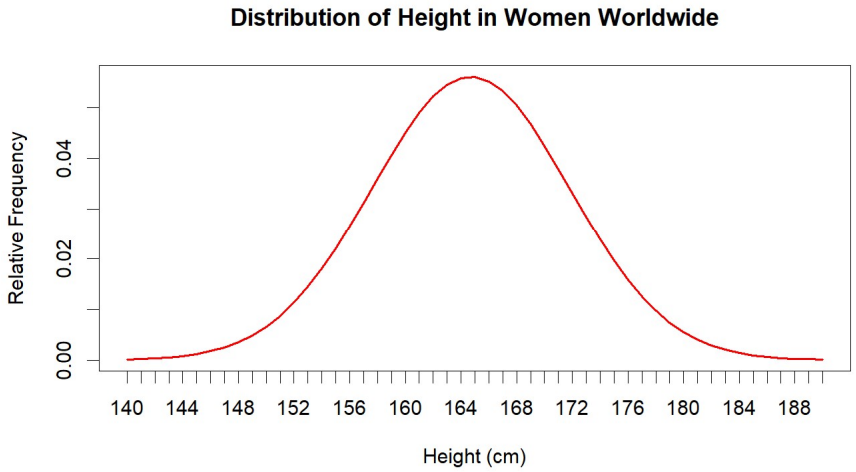
$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{X - \bar{X}}{s}$$

Example The heights of men and women around the world tend to follow a normal distribution. For women, $\mu = 164.7$ cm and $\sigma = 7.1$ cm. For a woman with $X = 175$ cm,

$$Z =$$

Example What is the Z-score of a woman who is 178.9 cm in height? What percentage of women have a Z-score below that?



Example If a person's standard score for IQ is $Z = 3$, what is their IQ? $(\mu = 100, \sigma = 15)$

In general, solving for X we find:

$$X = \mu + Z \cdot \sigma$$

$$X = \bar{X} + Z \cdot s$$

The Empirical Rule therefore says that the percentage of individuals with Z between:

-1	and	$+1$	is
-2	and	$+2$	is
-3	and	$+3$	is

Unusual Values

If X is a normally distributed variable, we consider any individual *unusual* if:

$$Z < -2 \quad \text{or} \quad Z > +2$$

Note: This means that about 5% of individuals are unusual for any given normal variable.

Example For adult men, the parameters for body height are: $\mu = 178.4$ cm and $\sigma = 7.6$ cm. The basketball player LeBron James is 203 cm in height. Is LeBron *unusual* in height?

What male heights would be considered unusual?

Chebyshev's Rule

If a variable X is *not* normal, then we cannot apply the 68-95-99.7 rule. Instead, we have only a *weaker* conclusion (named after the Russian mathematician Chebyshev) that says:

> For any variable X, the fraction of individuals with Z-score between $-k$ and $+k$ is at least: $1 - \frac{1}{k^2}$

Example What fraction of individuals must have a Z-score between -3 and $+3$ for *any* numerical variable (height, age, income, IQ, whatever...)