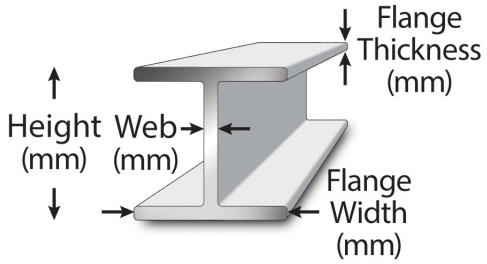
Set 1: Basic Terminology and Concepts

A **population** is a collection of objects of that we wish to study.

Example 1: Some examples of populations:

- All I-beams being made by a particular manufacturer.



- All Canadians who will be eligible to vote in the upcoming election.
- All Uvic students, staff and faculty.

In most cases, we cannot study a population directly for reasons of efficiency, practicality, or ethics. Instead, we select a **sample**, which is a subset of the population.

For example, we might select every tenth I-beam made by the manufacturing process, or take a random sample of 500 voters, or randomly select 40 Uvic students, staff and faculty.

We are usually interested in particular characteristics of the objects in the population. A **variable** is any characteristic that may vary from one member of the population to another.

The data that results from our observations may consist of one variable (a **univariate** data set), two variables (a **bivariate** data set), or many variables (a **multivariate** data set).

Example 2: If we measure the tensile strength of each I-beam, we have a univariate data set. If for each voter, we record the age and party of choice we will have a bivariate data set. If for each Uvic student, staff and faculty we record age, height, weight, and department, we will have a multivariate data set.

In a study, we are interested in some measurement of all of the population; this measurement is called a **parameter**. Some examples of parameters include: average value, median value, maximum value, proportion.

The measurement that we get from the sample that we actually study is called a **statistic**.

Example 3: We might be interested in the average yield strength of bars of a particular alloy of steel made in a certain foundry. In this case, the population is all bars of the alloy made in the foundry. The average yield strength would be the parameter of interest.

We take a random sample of 10 bars of the steel alloy, and find the average yield strength for those 10 bars is $593.2\ MPa$. This average yield strength is a measurement of our sample, so it is a statistic.

Usually, we use Greek letters to denote parameters and we use letters from the English alphabet to denote statistics. We will talk about this more in set 2.

In the branch of **descriptive statistics**, we organize, summarize, display, and describe features of the data.

Example 4: Some questions that descriptive statistics answers:

- What is the greatest tensile strength recorded? What is the range of recorded tensile strengths?
- What proportion of the sample of voters is older than 65?
- What is the average weight of the sample of people taking blood pressure medication? How spread out are the measurements for resting heart rate?

In the branch of **inferential statistics**, we try to draw conclusions about the population based on the measurements from the sample.

Example 5: Some questions that inferential statistics answers:

- What is a likely range of values of tensile strengths for all I-beams made by the manufacturer?
- Based on our survey, which party is likely to win the election?
- Can we conclude that there a relationship between weight and blood pressure?

Practice Question:

Determine whether the underlined words refer to a:

(A) Population

(B) Statistic

(C) Sample

(D) Parameter

- We wish to study poplar trees, so we make a selection of 15 poplar trees in a forest.
- From our selection of 15 poplar trees, we find the largest tree to have a height of 1.9 m.
- A newspaper wants to determine the feelings of Victoria residents regarding a bridge to the mainland.
- The newspaper phones 500 Victoria residents.
- It is found that 95% of these people are in favour of a bridge.

Set 2: Basic Descriptive Statistics

When we are collecting data, we are mostly dealing with samples.

The **sample size** is the number of observations. For a single sample, we denote the sample size by n.

If we have multiple samples, we use subscripts (e.g. n_1 is the size of the first sample, n_2 is the size of the second sample, and so on).

For now, we'll be considering a single variable x for a single sample of size n.

Individual observations will be denoted by $x_1, x_2, \ldots, x_n$. No sorting is assumed here.

Example 6: Suppose the following is data taken from some sample.

$$10, 6, 12, 7, 3, 6$$

$x_1 =$

$x_2 =$

$x_3 =$

$x_4 =$

$x_5 =$

$x_6 =$

Central Tendency: Mean, median, and mode:

The **mean** is another word for the **average** of a set of observations. $\overline{x}$ is the **sample mean** (mean of all observations from a sample). μ (pronounced as "mu") is the **population mean** (mean of all observations from a population).

Usually, μ is unknown (and unknowable). We use $\overline{x}$ as an *estimate* for μ.

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} x_i$$

Example 7: Suppose the following is data taken from some sample. Calculate the sample mean.

$$10, 6, 12, 7, 3, 6$$

Note: If we are given raw data, we can use the Sharp EL510*** calculator **STAT MODE** to find $\overline{x}$ easily.

The **median** is the "middle" observation of a *sorted* or *ordered* data set. We use $\widetilde{x}$ (x tilde) to represent the **sample median**, and use $\widetilde{\mu}$ to represent the **population median**.

If n is odd: $\widetilde{x}$ is the middle observation of the **sorted** set.

Example 8: Suppose we have the sample data: $6, 9, 3, 18, 11$. Find the sample median of these data.

If n is even: $\widetilde{x}$ is the average of the two middle observations of the **sorted** set.

Example 9: Suppose we have the sample data: $6, 9, 3, 18$. Find the sample median of these data.

Note: The Sharp EL510*** calculator cannot find the median of a data set; on a test you would need to find sample medians manually. Remember to ALWAYS **sort** the data before finding the median.

The median is **less** affected by **outliers** (extremely large or small observations). We can also say that the median is **insensitive** to outliers. The mean is **more** affected by outliers, or sensitive to outliers.

Example 10: Consider the following sets of sample data:

Set 1: $8, 5, 7, 6, 3$

Set 2: $800, 5, 7, 6, 3$

For both data sets, the sample median is 6. The first data set has a sample mean of 5.8, while the second has a sample mean of 164.2.

The **mode** is the observation that occurs the most frequently. The mode of $3, 5, 9, 9, 9, 5$ is 9, since 9 occurs the most often. There might be one mode, many modes, or no modes in a set of observations.

Example 11: The data set $1, 2, 3, 3, 3, 4, 4, 4, 5, 5$ has two modes (3 and 4).

The data set $1, 2, 3, 4, 5$ has no modes (since there is no observation that occurs more frequently than any other observation).

Data types: There are two broad categories of data types, **numerical** (e.g. heights in cm; mass in kg; a rating on a scale from 1 to 10, the percentage of people with a particular genetic marker) and **categorical** (e.g. a plant's species; a person's name; an answer of "yes" or "no" to a survey question).

While we can find the mode for any kind of data, the median and mean can only be calculated from numerical data.

Caution: Often times, we record categorical data as numerical values. For example, when collecting data about colours, we may record "blue" as 1, "red" as 2, etc. While we can compute mean, median, and standard deviation (will be defined shortly) of this type of data set, there is in general no meanings to these values.

Variability:

s^2 and σ^2 (pronounced as "sigma square") denote the **sample variance** and the **population variance**, respectively. They are measures of the amount of variation or dispersion of the data set.
A low variance indicates that the values tend to be close to the mean of the data set, In other words, the values are less spread out or more consistent. While a high variance indicates that the values are spread out over a wider range.

The i^{th} **deviation** is the difference between x_i and $\overline{x}$. A useful fact about deviations:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

The **sample variance**, denoted s^2 is:

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

s and σ are the sample and population **standard deviation**, respectively. As the symbols imply, the standard deviation is just the square root of the variance. It also has the same unit as the measurement values.

A computational (shortcut) formula exists to assist with hand calculation of sample variance. It uses the fact that

$$\sum (x_i - \overline{x})^2 = \sum x_i^2 - n(\overline{x})^2$$

so,

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

□

Example 12: Find the variance and standard deviation of the following sample:

$$7, 7, 9, 15, 16, 17, 19, 21, 22, 40$$

Note: If we are given raw data, we can use our calculator **STAT MODE** to find s easily.

The **coefficient of variation** (cv) is a *dimensionless* quantity (i.e. no units of measurement) which can be used to compare the variability of other sets of observations with different numerical measures.

The cv is calculated by $\frac{s}{\overline{x}}$

The greater the cv, the more spread out, i.e., more varied, are the data.

Example 13: One set of observations has a mean of 35 with a standard deviation of 7. A second set of observations has a mean of 55 with a standard deviation of 9. Which data set has more variability?

Because cv is dimensionless (i.e. it doesn't have a unit of measurement), it is not affected by the choice of measurement units for the data.

Example 14: Jem and Kimber each measure the heights of the same 5 people. Jem uses m as her units, and Kimber uses cm.

Jem's Data: $1.71, 1.73, 1.81, 1.82, 1.80$

Kimber's Data: $171, 173, 181, 182, 180$

The standard deviation of Jem's measurements is approximately $0.0503\ m$, while for Kimber it is $5.03\ cm$.

However, for both Jem and Kimber, the cv is the same: approximately 0.02835

Suppose we are summarizing a data set. When might we choose to use the **median** rather than the mean as a measure of the "centre"?

(A) When you suspect there are no outliers in the data set.

(B) Whenever you feel like it – either is a appropriate to use at any time.

(C) When you suspect there are outliers in the data set.

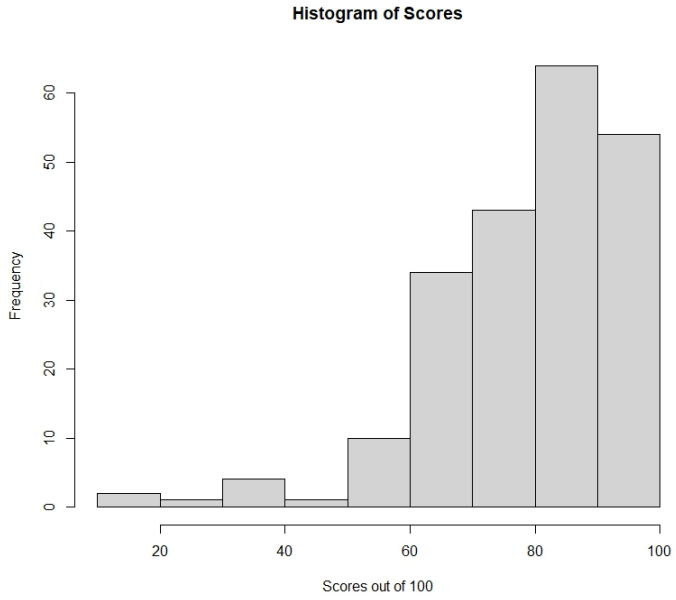
Suppose we have two data sets. The first data set has a standard deviation $s_1 = 5$ and the second data set has a standard deviation of $s_2 = 10$.

(A) The first data set is more "spread out" than the second data set.

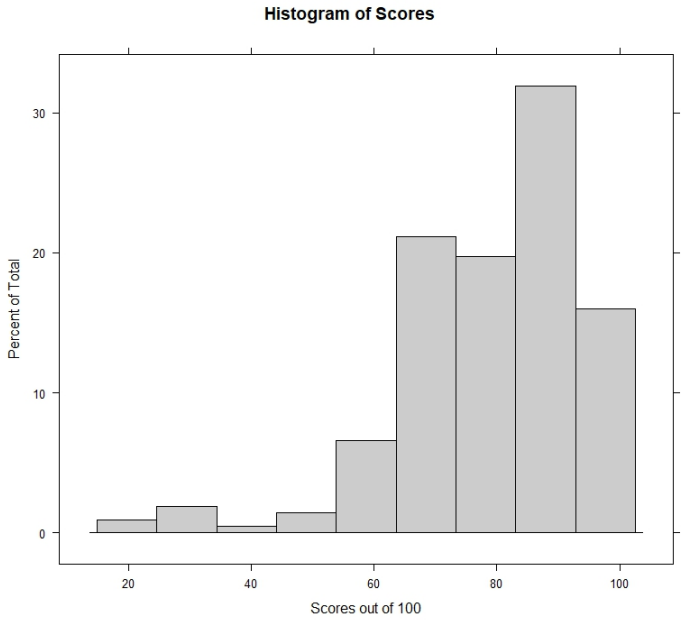
(B) The second data set is more "spread out" than the first data set.

(C) We don't have enough information to tell which data set is more "spread out".

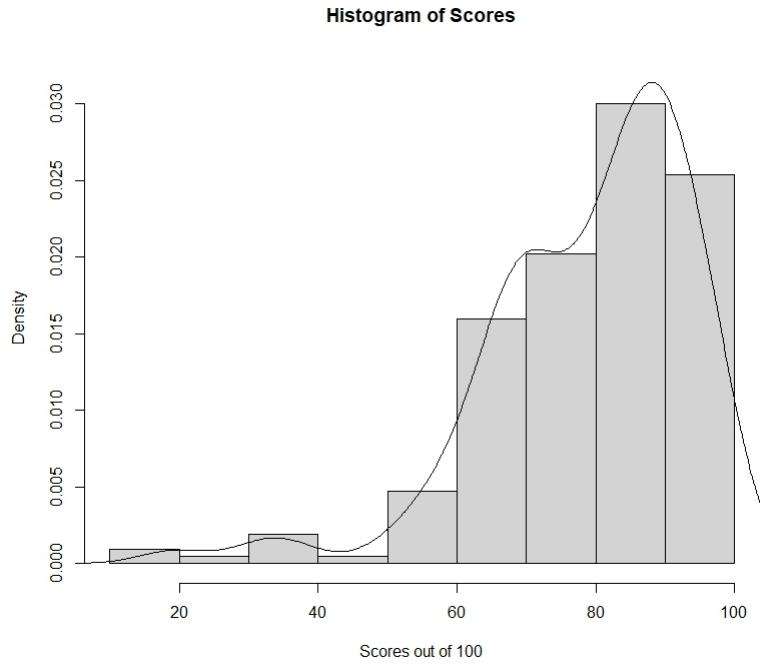
A **histogram** is a visual representation of a frequency distribution. The horizontal axis is for our measurements, and the vertical axis is for the frequencies. The following is a histogram of 213 final scores (out of 100 points).



Alternatively, we may use the vertical axis for the relative frequencies. The relative frequency is the proportion of times the value occurs.



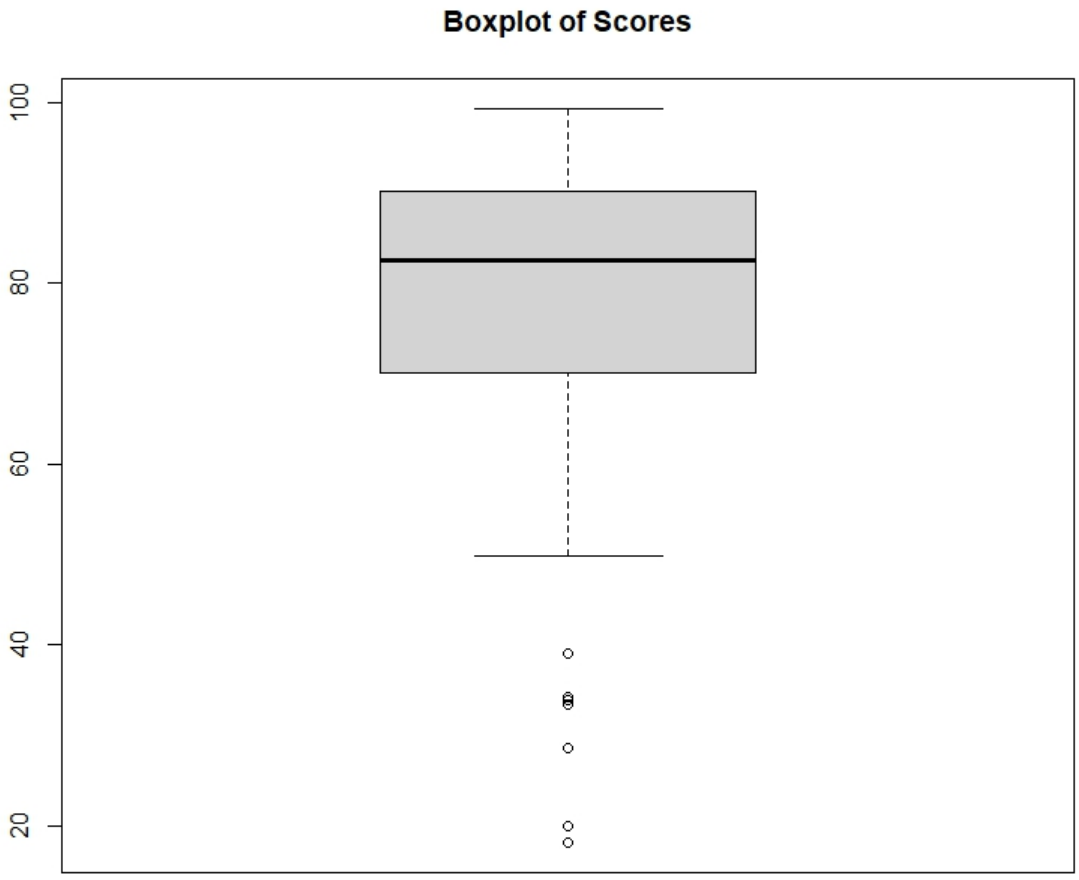
We can imagine our histogram as being approximated by a smooth curve. We can classify our frequency distribution by the shape of these curves. **Note:** The area under this density curve is 1.



Curves may be **symmetric** or **asymmetric**. They may have one peak (**unimodal**), two peaks (**bimodal**), or many peaks (**multimodal**). For unimodal distributions, they may be **positively skewed** (the right tail is "stretched") or **negatively skewed** (the left tail is "stretched").

We would describe the histogram of the marks as being asymmetric, unimodal, and negatively skewed.

Another useful visual summary is a **boxplot** or **box-and-whisker plot**. The boxplot allows the viewer to visually identify the median, the **lower quartile** (the median of the lowest half of the observations), and the **upper quartile** (the median of the upper half of the observations).



The **interquartile range** (IQR) is the distance from the lower quartile to the upper quartile (i.e. the length of the box in the plot).

Outliers: The IQR is commonly used to identify outliers. A typical procedure is to consider any observations outside of the interval [lower quartile - 1.5 x IQR, upper quartile + 1.5 x IQR] to be an outlier.

Example 15: Suppose we have the following sample data:

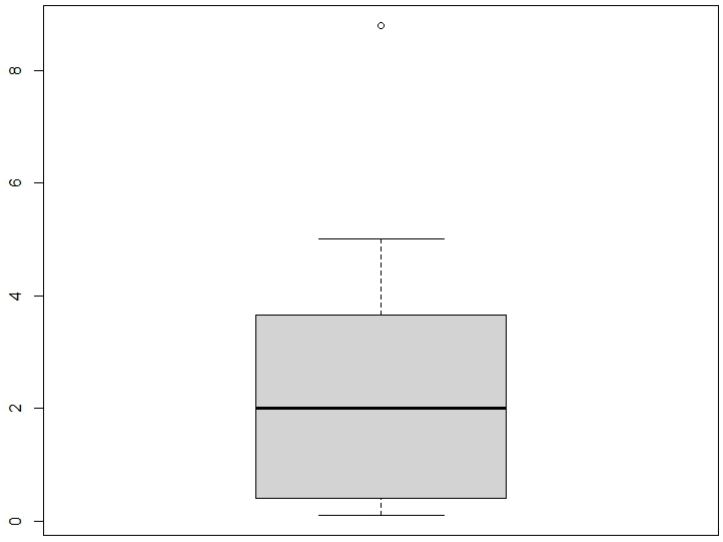
1.7, 0.9, 3.8, 2.1, 1.9, 0.6, 0.5, 5.0, 2.4, 0.1, 5.0, 0.3, 8.8, 0.3, 0.3, 3.3, 4.8, 0.2, 2.2, 3.5

We've used R to find that the lower quartile is 0.45, the upper quartile is 3.575, and the IQR is $3.575 - 0.45 = 3.125$

Any observations outside of the interval

$$[0.45 - (1.5)(3.125),\ 3.575 + (1.5)(3.125)] = [-4.2375, 8.2625]$$

would be an outlier. If we looked at our data, we would find one observation (8.8) which is outside that range. The top whisker ends at 5.0 (our largest non-outlier), and the outlier 8.8 is indicated with a circle.



Set 3: Correlation Coefficient

Bivariate data: data with two variables. A bivariate data set will be a set of ordered pairs: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

A frequent question we ask with bivariate data is whether or not there is a relationship between the two variables.

A **scatterplot** is used to visually depict bivariate data. The observations are plotted as a set of points on the plane.

Important: For a scatterplot to be appropriate, each pair of measurements must be made on the same object.

Example 16: We select 20 people, and for each person, we record x, their age, and y, their maximum heart rate.

Here, the data is clearly bivariate (one sample of size $n = 20$, with pairs of measurements being made); a scatterplot would be appropriate.

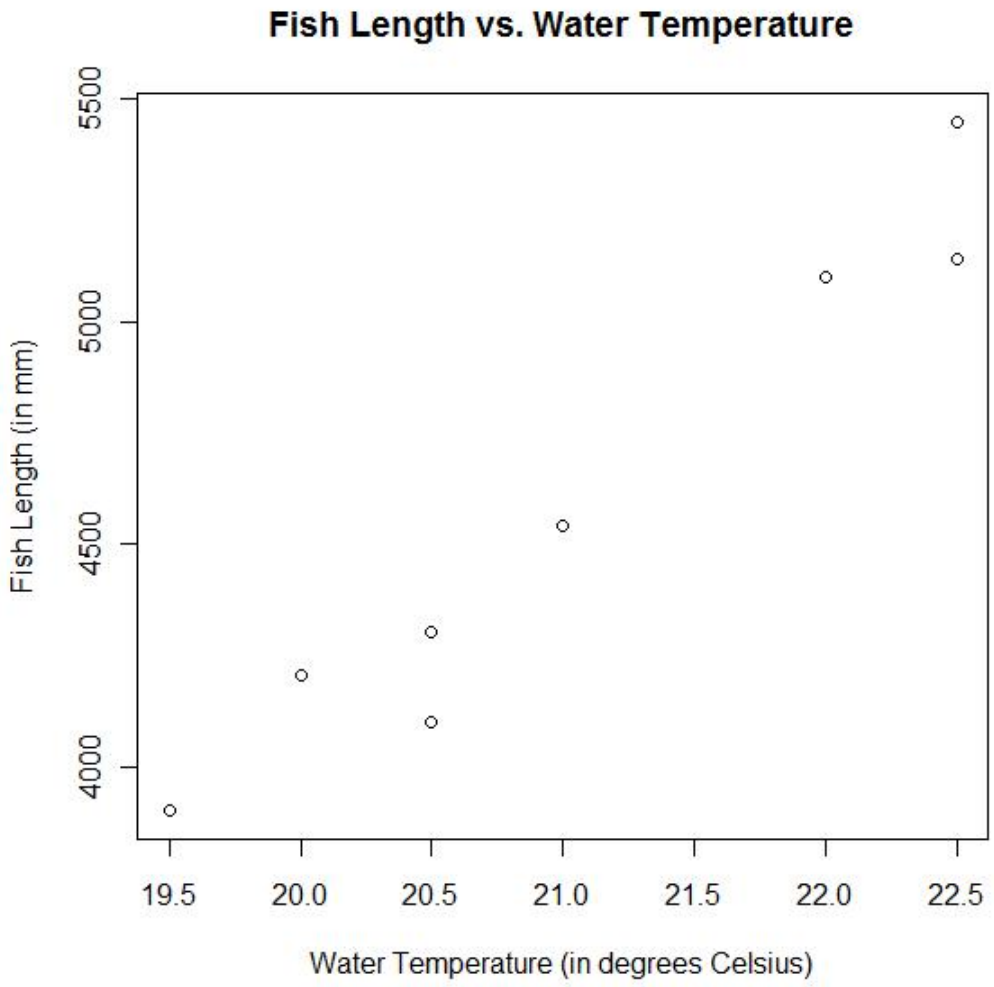
Example 17: We select 20 people and put them on Diet A, and measure x, their blood pressure after two weeks. We select another 20 people and put them on Diet B, and measure y, their blood pressure after two weeks.

Here, we have two samples, of sizes $n_1 = 20$ and $n_2 = 20$. The data is **not** bivariate; a scatterplot would **NOT** be appropriate.

Example 18: Several of a particular species of fish are grown from eggs in tanks set at particular temperatures. After a fixed number of days, all fish are measured.

We wish to investigate the relationship between y, the length of the fish (in mm), and x, the temperature of the tank (in degrees Celsius).

x	19.5	20	20.5	20.5	21	22	22.5	22.5
y	3900	4205	4100	4300	4540	5100	5450	5140



The **sample correlation coefficient**, denoted r, can be used to assess the **linearity** of bivariate data.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

The computational form is a little faster to use:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}$$

We might recognize that the denominator could be written in terms of s_x and s_y (the standard deviation of x and y, respectively).

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{(n-1)s_x s_y}$$

Yet another formula for computing r looks like this:

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{(n-1)}$$

is called the sample covariance of the 2 variables in the data set.

Example 19: For our fish data:

We find $\sum_{i=1}^{n} x_i y_i = (19.5)(3900) + \cdots + (22.5)(5140) = 778165$.

Then, we find $\overline{x}, \overline{y}, s_x, s_y$ using our calculator.

We have $r \approx 0.973$.

Interpretation: r (no unit), takes on values between -1 and 1, inclusive.

An r value of -1 indicates a perfect negative linear relationship.

An r value of $+1$ indicates a perfect positive linear relationship.

An r value of 0 indicates no **linear** relationship.

If r of a bivariate data set is $+1$, then a straight line with a positive (but unknown at this point) slope can be drawn that passes through all points on the graph.

Similarly, if r of a bivariate data set is -1, then a straight line with a negative (but unknown at this point) slope can be drawn that passes through all points on the graph.

Example 20: For the fish data, the correlation coefficient is positive (indicating an increasing relationship), and is close to 1. So the linear relationship between fish length and water temperature is quite strong.

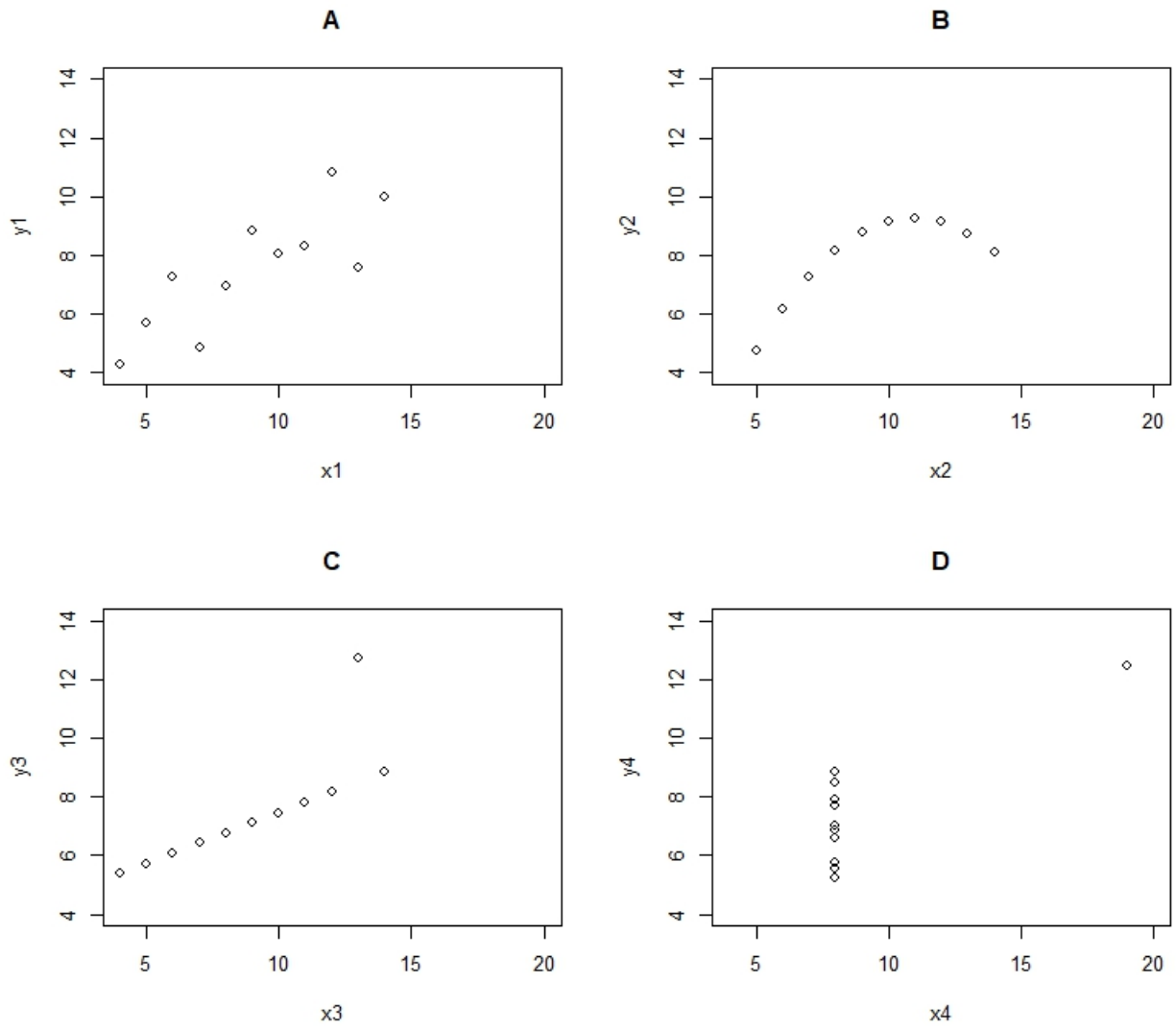
Note: r is **insensitive** to positive scaling and shifting, that is, if you multiply a variable (x and/or y) by a positive factor, or add a constant to a variable, r doesn't change.

Warning: An r value of 0 does not mean there is no relationship, only that the relationship is not linear.

Example 21: Consider the following observations:
$(-3, 9), (-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4), (3, 9)$

These data have an r value of 0. There is no linear relationship. However, if you examine a scatterplot there is a clear *quadratic* relationship.

Example 22: Which data set has the highest r value?

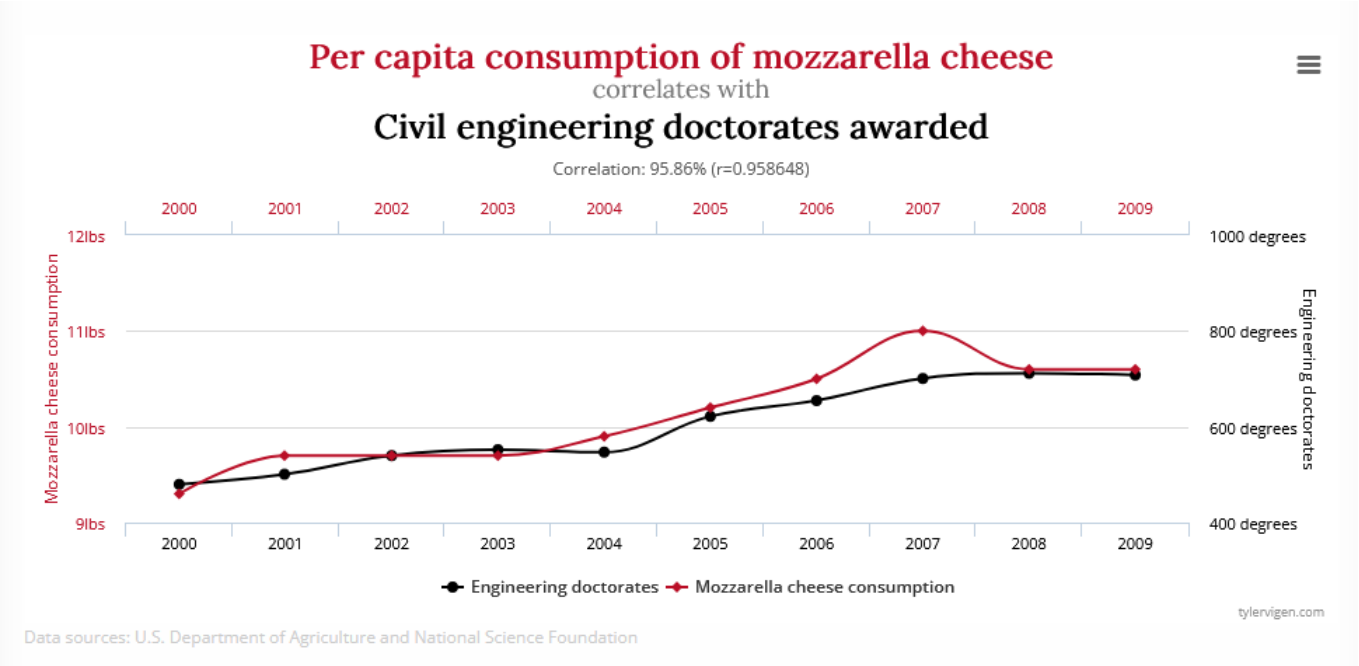


A B C D

Important: When we examine variables x and y and find there appears to be some correlation between them, there are many possible explanations:

1. x causes y.
2. y causes x.
3. There is (are) some other unexplored variable(s) which relates to both x and y.
4. The correlation is spurious (there's no actual relationship; the correlation is just a coincidence).

Example 23: The image below shows that there appears to be a very strong correlation between the per capita consumption of mozzarella cheese and the number of civil engineering doctorates awarded. This is one of many examples of spurious correlation (or is it?).



(Image from `tylervigen.com`)