**Statistics**

**Types of Data:**

**Primary Data** is gathered by or for the person who is going to use it.

**Secondary Data** is not gathered by the person who is using it.

**Numerical** - Any data answered with numbers:

* **Discrete** - Data with exact values
* **Continuous** - Data without exact values

**Categorical** - Any data not answered with numbers:

* **Ordinal** - Data that can be ordered
* **Nominal** - Data cannot be ordered

**Finding, collecting and organising data**

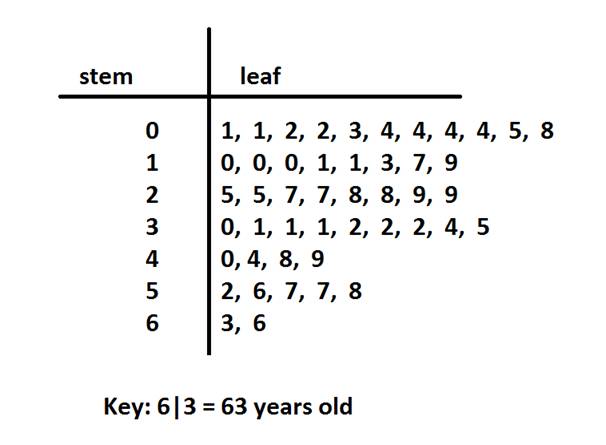
* **Population** - The entire group being studied
* **Census** - A survey of the entire population
* **Sample** - A subgroup selected from the population
* **Parameter** - A variable/measurement of some characteristic of the population.
* **Statistic** - An estimate of some parameter, estimated through our study.
* **Simple Random Samples** are the commonly used method for finding a sample. This means choosing people randomly from the population, with everyone having an equal chance of being chosen.
* **Importance of representativeness**: A representative sample accurately represents the population being studied. It helps to avoid biased samples, which means we avoid biased results.
* **Sample surveys** mean collecting data from a sample the population using a survey/questionnaire.
* **Observational studies** mean observing and recording data without any intervention by the people doing the study.
* **Designed experiments** mean manipulating/designing **explanatory variables**(like dosages of a drug taken in medial trials)in our data to observe and record the effect on **response variables**(effects of our drug).

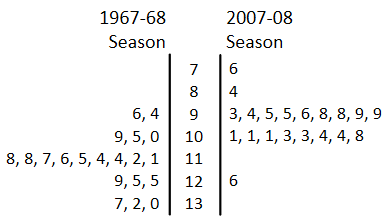
When making a plan to gather statistics/data we must then:

* Find define the population being studied and take a representative sample.
* Decide on the type of study (survey, observational, or designed experiment).
* Decide on how to collect the data, through as a survey or observation checklist etc.
* Collect data from the sample and record it.
* Analyse and interpret the data to draw conclusions.

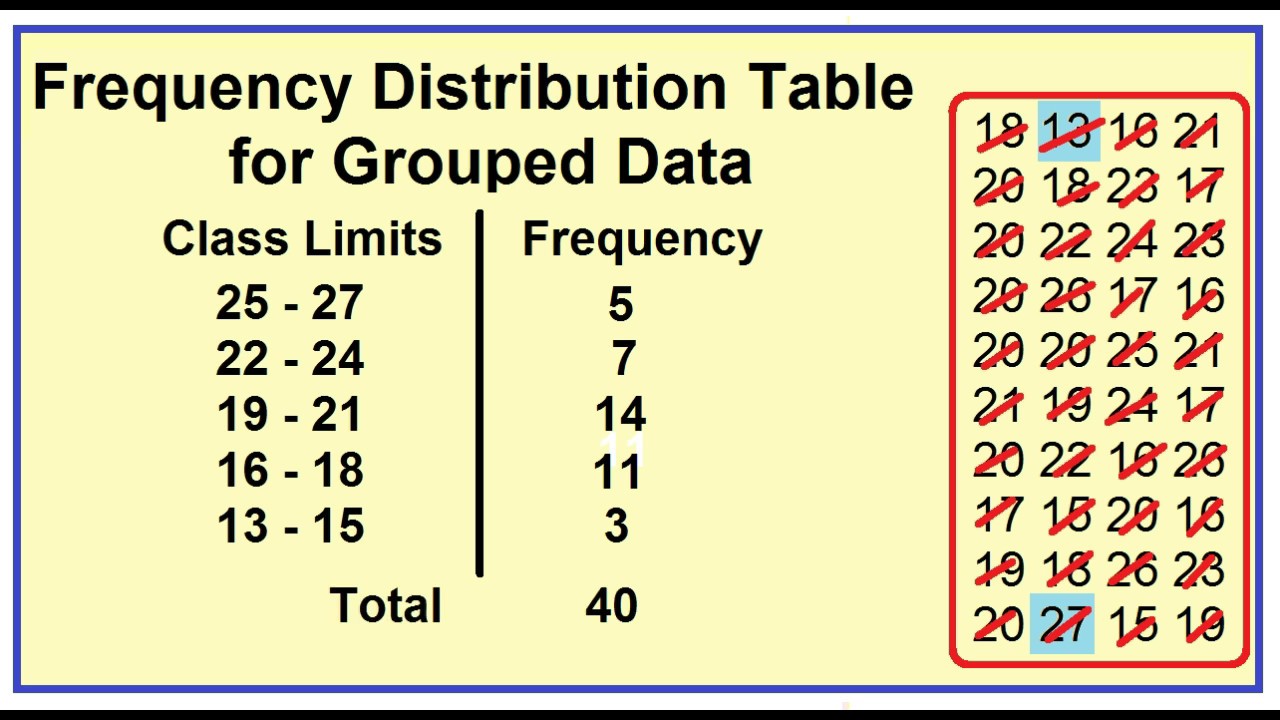
**Graphing Data**

Steam & leaf plots:

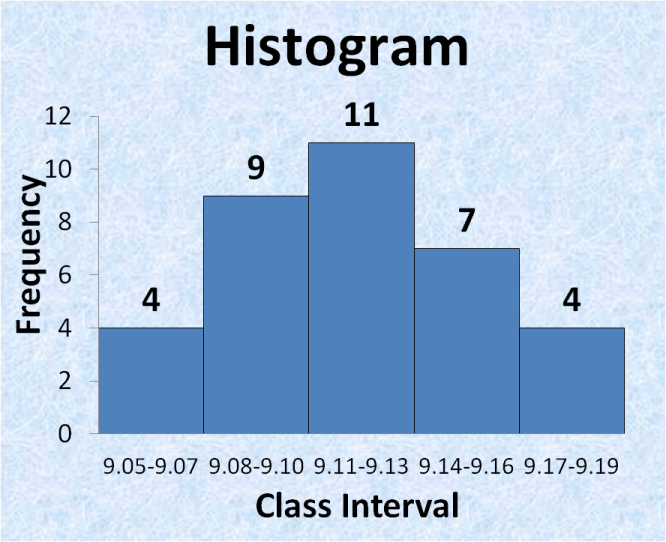


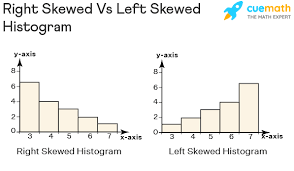


Frequency Tables

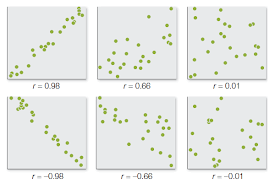


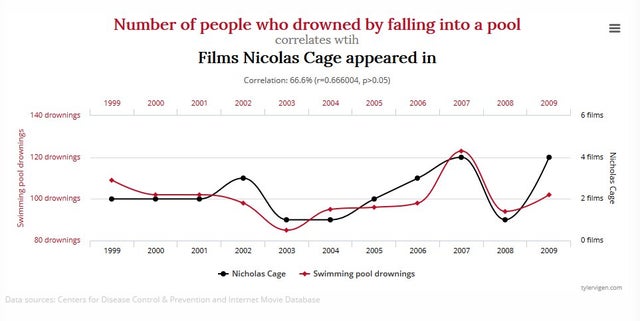
Histograms:





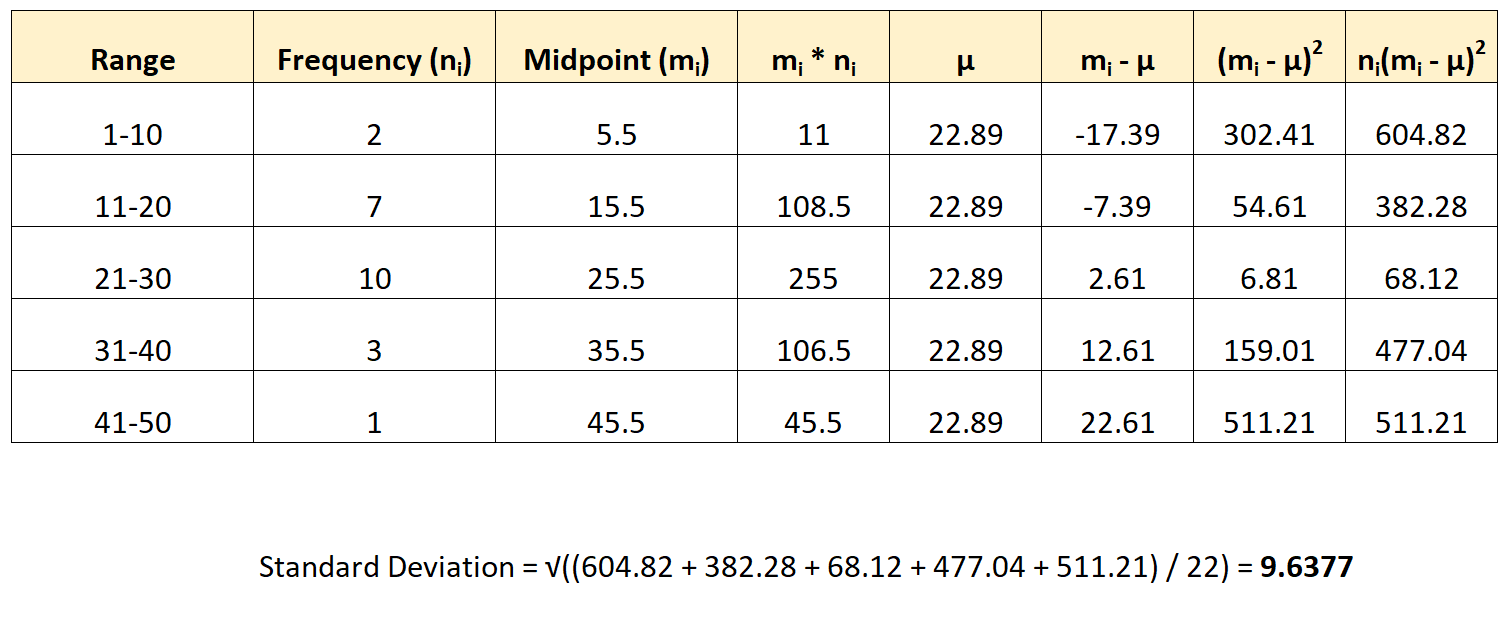
Scatter Plots



Correlation Doesn’t Equal Causation:

**Analysing Data:**

* **Mean**: The sum of the values over the number of values
* **Median**: The middle value of all values. If there is no middle, we find the mean of the middle two.
* **Mode**: The value that occurs the most
* **Range**: The difference of the lowest and highest values
* **Interquartile Range**: The Range of the of middle half of the values. The highest and lowest quarters are written off as outliers.
* **Standard Deviation**: The measure of how far our values are from the mean. This can be found by finding the average of the squares of the distance from each data point to the mean. This gives us the **variance**. The square root of the variance is the standard deviation. This frequency table shows how the standard deviation is calculated with ranges of values.



* **Percentiles**: Values that divide the data into 100 equal parts.

c = (k x n)/100

where c is our value

k is our percentile

n is the number of values

If there is no exact value, we take the average of the numbers above and below, the same way we do with the median(The median is the 50th percentile).

* **Sampling Variability:** If multiple samples are taken from the same population, they will differ from one another. This means that information obtained from a single sample may not be representative of the entire population. The size of the sample and the variability in the population can affect how much sampling variability there is.
* **Measures of central tendency** (mean, median, mode) and **measures of dispersion** (range, variance, standard deviation) can be used to describe variability in samples.
* **Confidence intervals** and **hypothesis testing** can be used to draw inferences about the population based on the sample.
* **The Empirical Rule**: For a normal distribution, approximately **68%** of the data falls within one standard deviation of the mean, **95%** falls within two standard deviations, and **99.7%** falls within three standard deviations. This can be used to make decisions about the likelihood of an event occurring based on the distribution of data.

**Hypothesis Testing:**

* **Hypothesis Test**: Determining whether a hypothesis about a population is supported by the sample data.
* **The null hypothesis:** The hypothesis that there is no significant difference between the sample and the population.
* **The alternative hypothesis**: The hypothesis that there is a significant difference.
* **Population Proportion**: The proportion of the population that some statistic applies to. Represented by **p**.
* **Sample Proportion**: The proportion of the population that some statistic applies to. Represented by **p̂**.
* The  **margin for error**(**E**) = , √n where n = the number of people in the sample. This is used in measuring accuracy of a sample.
* The **standard error** is the amount by which the **sample proportion** may differ from the **population proportion**. To calculate the standard error (**SE**) for a population proportion, you need to use the formula:

Where Z is the z-score for the level of confidence (e.g. for 95% confidence, Z=1.96)

p is the estimated proportion of the population(if known)

n is the sample size.

* Once you have calculated the standard error, you can conduct a hypothesis test on a population proportion using the margin of error. This involves comparing the observed proportion in your sample to the hypothesized proportion in your null hypothesis.
* The null hypothesis (H0) is the default assumption that there is no difference between the sample proportion and the population proportion, while the alternative hypothesis (Ha) states that there is a difference.
* **To conduct a hypothesis test** for a population proportion, you can use the following steps:
* State the **null hypothesis** and **alternative hypothesis**.
* Determine the level of significance (alpha) for the test.
* Calculate the standard error, and check whether our null hypothesis lies between the standard error on our sample proportion.