* **Structured vs. unstructured data**

**Structured data is in tabular form (organized in rows and columns). Strict structure - can be read and used by a computer. Typically stored in relational databases.**

**Unstructured data has no pre-defined structure (images, videos, sound files, and texts like email, social media posts, or ratings). More pre-processing necessary before data can be analysed. Typically stored in document databases.**

* **Qualitative / Categorical Data Type**

**Qualitative data describes observable categories (colour of eyes). Calculating statistics and analysing qualitative data is more restricted. E.g., Can calculate average length, but average colour does not make sense.**

**Nominal - Assigns names to each data point without placing it in order**

Employment status

Soccer player position

Results of a test could be classified nominally as a "pass" or "fail"

**Ordinal - Groups data according to a ranking and ordering system**

CH satisfaction – unsatisfied, neutral, satisfied

Income status – low, middle, high

Results of a test could be grouped in descending order by grade: A, B, C, D, E and F

* **Quantitative / Countable Data Type**

**Quantitative data describes something numerically; it can be measured or counted (distance between home and work).**

**Discrete - Interval / Count - Numerical values which fall under integers or whole numbers**

# of trees in a field

# of t-shirts sold in a store

**Continuous – Numerical values which fall under floats or fractional numbers**

Race car lap time

Height in cm

* **Data uses**

**Descriptive** – Quantitatively describe results with all data

**Exploratory** – Looking for previously unknown relationships

**Inferential** - Infer results / test theories with sample data

**Predictive** – Analyse current data to predict future trends / events

* **Population** - An entire collection of observations or events
* **Sample** - A subset from a population meant to represent the population
* **Random Sampling**  - Generate random numbers column and select samples
* **RAND () - Random between 0 - 1**
* **COUNT ()**
* **POWER (a, b) or ^**
* **SQRT**
* **AVERAGE - Mean - *Affected by outliers***

Sum of samples / # of samples

* **MEDIAN - Median - *Robust to outliers***

Middle value, sorting smallest to largest (mean of middle 2 numbers if even)

50% lower than median, 50% higher than median

* **MODE - Mode - *Best for categorical data***

Number that appears most often (most frequent) in a dataset

* **Symmetrical Distribution = Normal Distribution = Bell Curve -** Mean and median align well when distribution is symmetrical
* **Asymmetrical Distribution -** Occurs with outliers - ***Use mode to describe typical value***
* **Empirical Relationship - Mean – Mode = 3 (Mean – Median)**
* **Spread / Range -** How far apart data points are - variety of data (= maximum – minimum)
* **Distance** - Dispersal between each data point to the mean (= value – mean)
* **VARP - Variance - Dispersal of dataset from mean** (

= sum of all distance 2 (squared) / total number

***Small variance = data is less spread***

***Large variance = data is more spread***

1. Calculate mean
2. Distance - Subtract mean from each value
3. Square differences from the mean
4. Take average of squared differences

* **STDEVP - Standard Deviation for population - Put variance in scale of original data**

Standard deviation =

1. Take the square root of the variance by SQRT (variance) or STDEVP

***Standard Deviation close to 0 = data clustered around the mean***

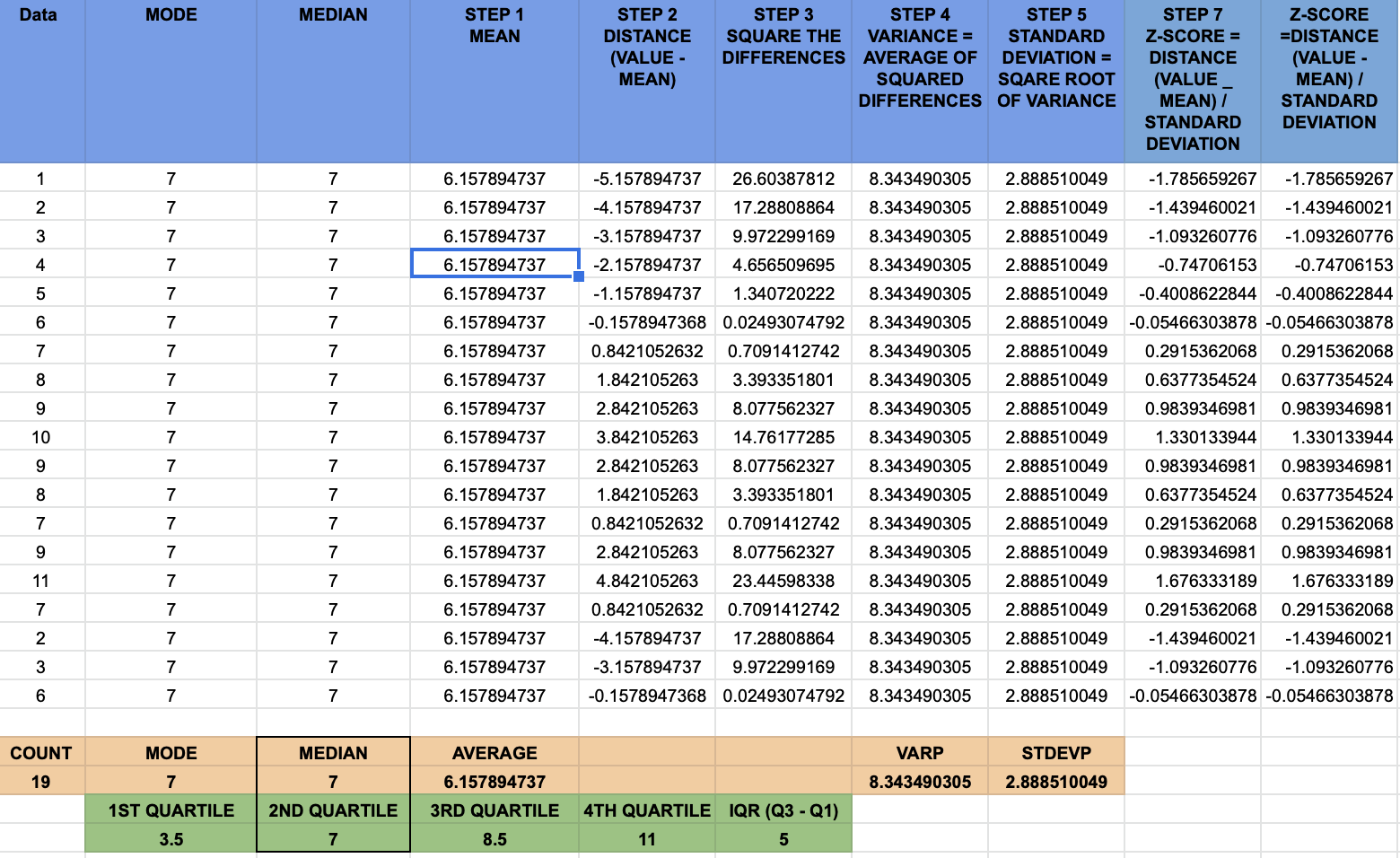
***Can be determined by looking at the spread in a histogram (x-value range)***

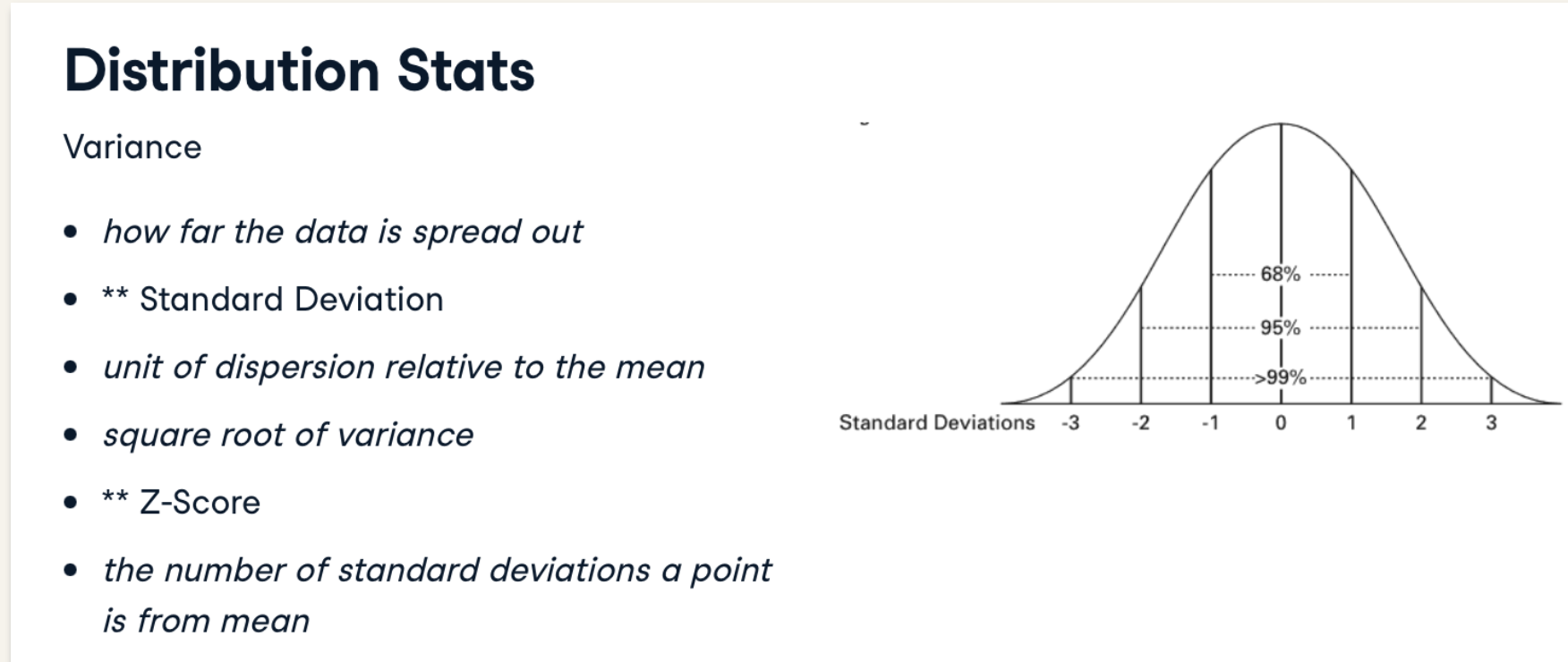
* **STDEV - Standard Deviation for sample**
* **STANDARDIZE (DATA POINT, AVERAGE, STDEVP) - Z-Score (Standard Scores)**

Centres distribution around the mean and calculates the number of standard deviations away from the mean each point is

= (data point - mean) / standard deviation

1. subtract mean from each data point
2. divide result by standard deviation

****

******

* **INSERT > CHART > HISTOGRAM**

Visualises a single variable

Al least over 20+

Correctly set bins

* **INSERT > CHART > SCATTER CHART - Scatter Plot / XY Graph**

Visualises relationship between 2 variables

Each point represents row in data

Points are scattered along coordinates

Add Trendline

* **LINEST (range y, range x) - Line Estimate**

Slope and intercept of trend line - returns slope, then intercept

Intercept is the point where line intersects Y axis

Slope declares trend relationship between the variables

Order of x and y tables matters

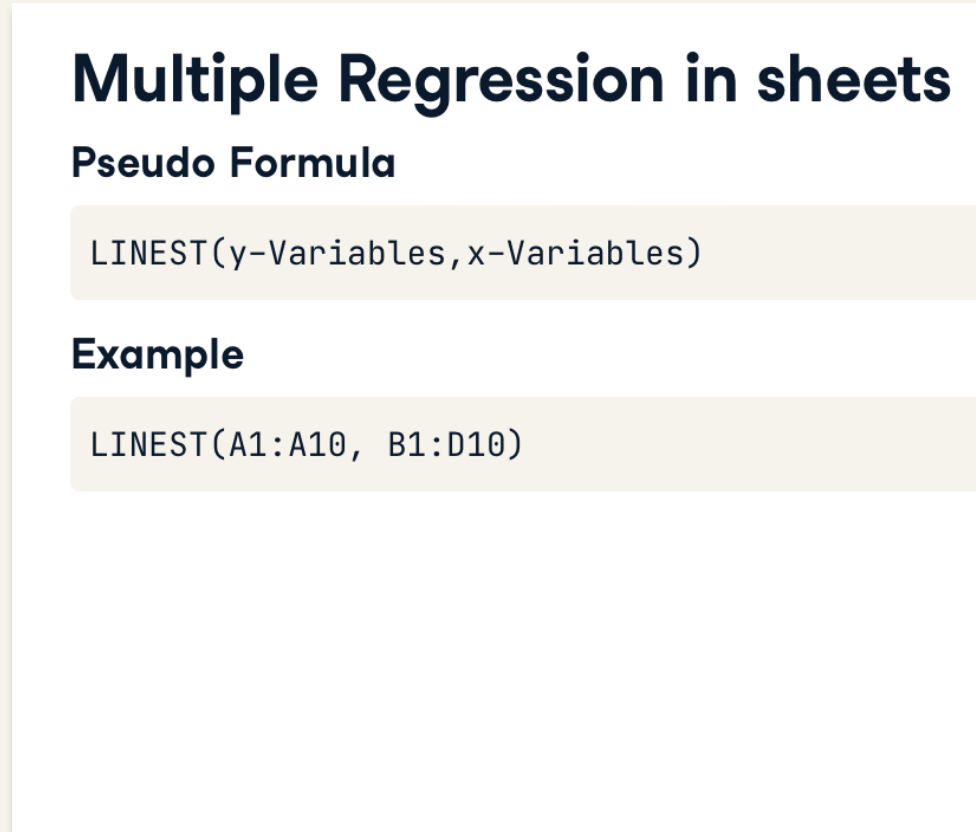
* **FIT = Intercept + (slope \* data point)**
* **Linear Regression**

Predict value of a variable based on value of another variable

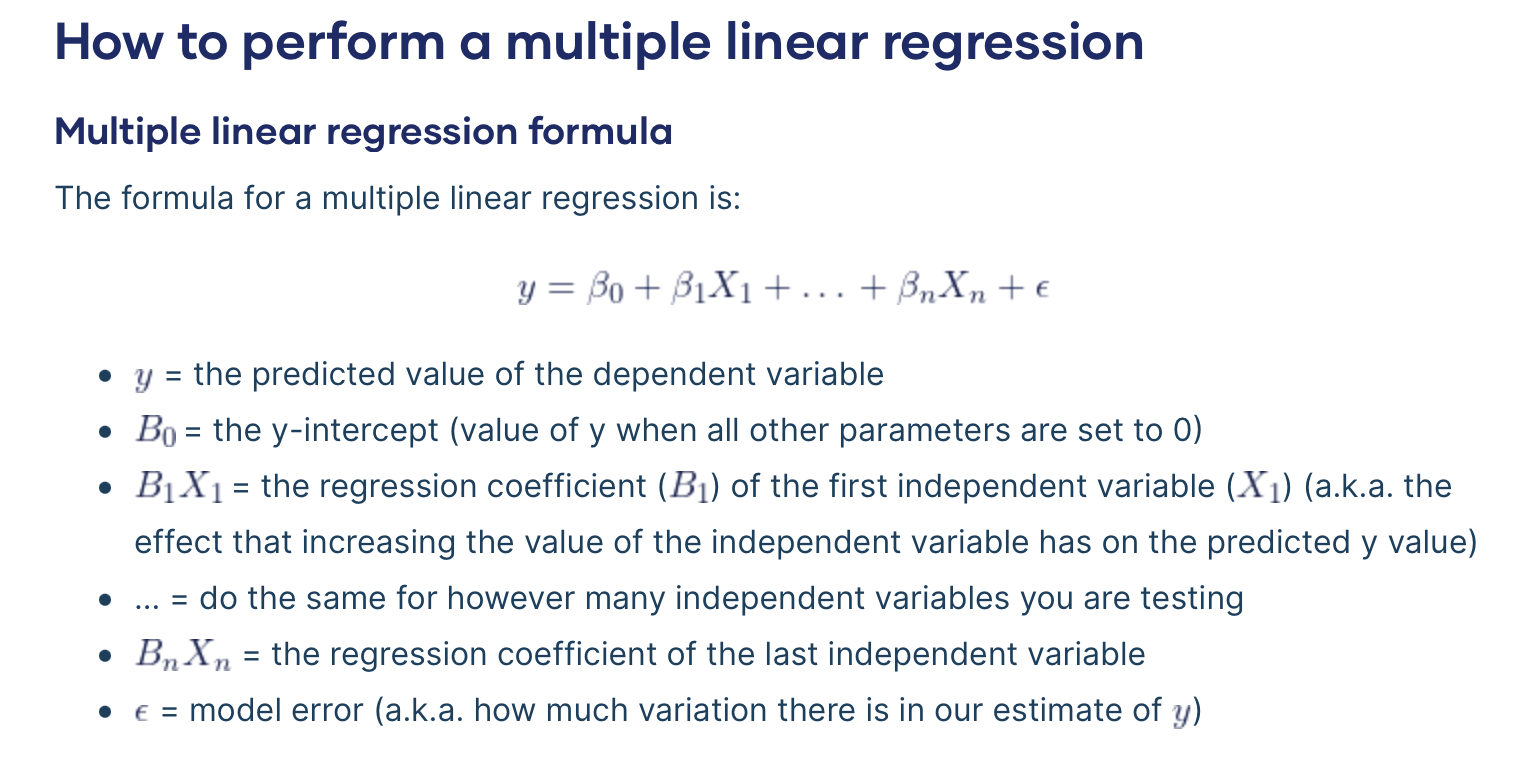
Variable predicted - dependent variable

Variable used to predict - independent variable

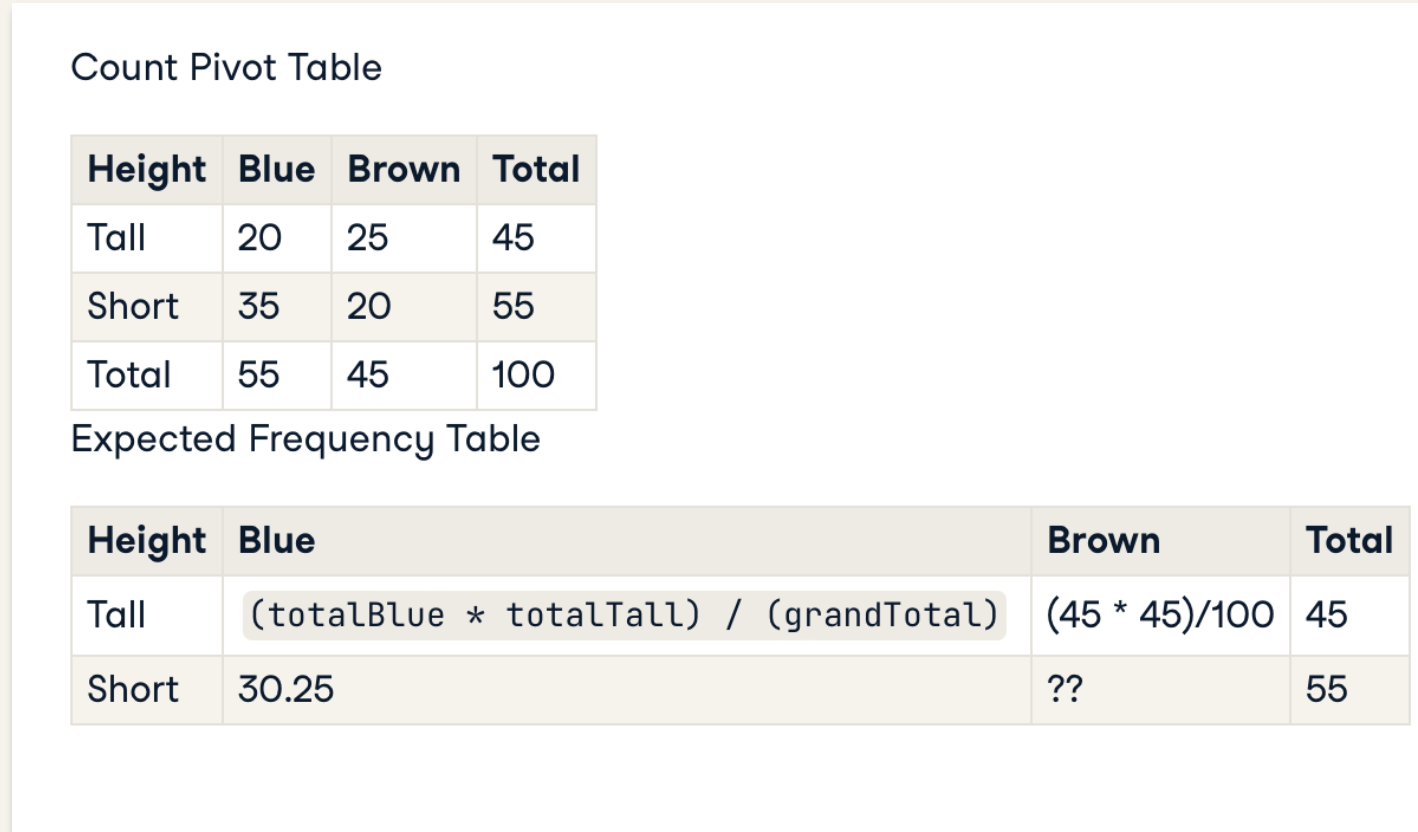
* **Multiple Linear Regression**



* **Multiple Linear Regression Formula**

****

* **Expected Frequency**

****

* **CORREL - Correlation**

Ranges between -1 and 1

0 correlation means there is no relationship between variables

Positive correlation indicates that as one variable increases, the other also increases

Negative correlation values signify that as one variable increases, the other decreases

* **INSERT > CHART > BAR CHART**

Visualises relationship between categorical data (distinct classes or factors)

Stacked

Side-by-Side

Proportional

* **COUNTIF (range, value)**

***Subtract from total if counting 0***

* **QUARTILE - Quartiles - Split data into 4 equal parts 0%, 25%, 50%, 75% & 100**

For each quartile, the value represents the percentage of values that are less than or equal to that number

***Second quartile (50%) = median (middle value)***

* **Interquartile Range (IQR) - How the middle 50% differs**

IQR = 3rd Quartile – 1st Quartile

***Less affected by extreme values, compared to standard deviation***

* **Box plots** – used to visualise quartiles

Left edge = 1st Quartile

Middle line = Median

Right edge = 3rd Quartile

***Extreme values (Outliers) = shown beyond the horizontal lines***

* **Probability**

P (event) = P (heads) = = ½ = 50%

***Always between 0% (impossible) - 100% (certain)***

* **Independent probability** - 2 events are independent if probability of 2nd event is not dependent on 1st
* **Sampling with replacement -** Select random sample and use again
* **Conditional Probability -** 2 events are dependent if probability of one event is dependent (conditional on) the outcome of another
* **Sampling without replacement -** Select random sample, but don’t use same sample again
* **Venn Diagrams**

**Diagram, venn diagram

Description automatically generated**

***Overlap will change depending on results of first event***

**E.g. Chance of order being worth more than $150, given it is for kitchen products**

= 20/161 DIV BY 161 / 1761

There are 581 orders > $150 and 161 orders for kitchen products

20 kitchen product orders are > $1

**E.g. Chance of order being for kitchen products, given it is worth more than $150**

**=** 20 / 1761 DIV BY 581 / 1767 = 3.4%

**Chart, bubble chart

Description automatically generated**

To calculate the conditional probability, d

* **Conditional probability formula = Probability of event A, given event B, is equal to the sum of Probability of event A and event B (overlap) Divided by the Probability of event B**

**P (A|B) =**

* **Discrete Distributions -** Situations with discrete outcomes: count or interval (counting dots for die rolls)
* **Expected Value = Mean of probability distribution**

Multiply each value by its probability and add together

**E.g. Expected value of a fair die roll =**

(1 x 1/6) + (2 x 1/6) + (3 x 1/6) + (4 x 1/6) + (5 x 1/6) + (6 x 1/6) = 3.5

* **Why are probability distributions important?**

Help to quantify risk and informed decision making

Hypothesis testing - Determine probability that results may have occurred by chance

Chart, histogram

Description automatically generated

***Each bar represents an outcome, and each bar’s height represents the probability of that outcome***

* **Probability = area**

Calculate probabilities of outcomes by areas of probability distribution in histogram

**E.g. What's the probability that our die roll is less than or equal to two?**

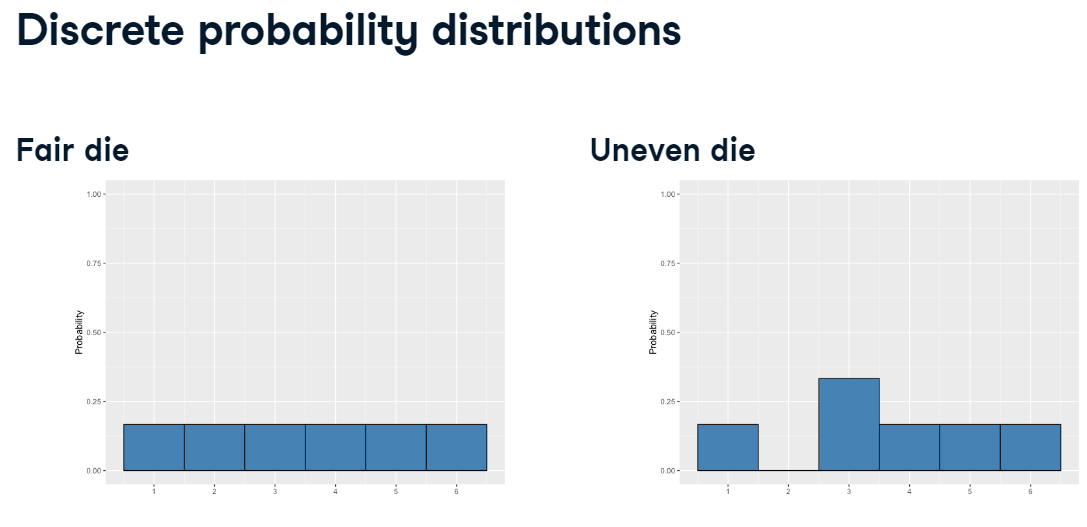
Chart

Description automatically generated

* **Uneven Die – 2 got turned to 3**

**Application

Description automatically generated with medium confidence**



* **Discrete uniform distribution**

When all outcomes have the same probability (e.g. fair die)

* **Law of large numbers**

Despite same probability for rolling each number, there are different results as the sample was random

As the size of sample increases, the sample mean will approach the theoretical mean (expected value)

Does not always result in a sample taking the form of a normal distribution as the sample size increases

* **Continuous distributions - -** Situations with continuous data

Bus arrives every 12 minutes, so if we show up at a random time, we could wait anywhere from 0 up to 12 minutes

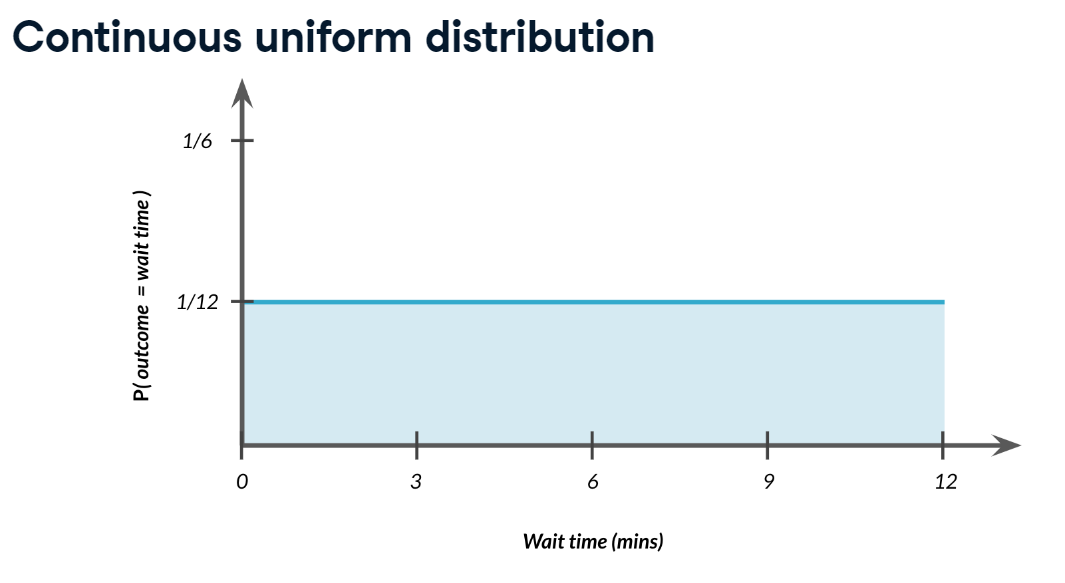
Infinite number of minutes we could wait

Therefore, we can't create individual blocks like we could with count or interval data.

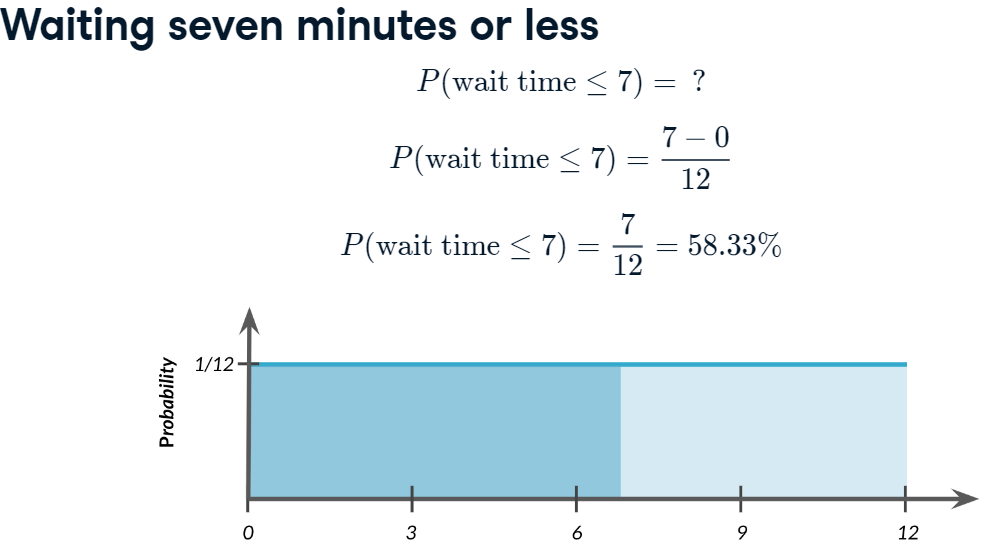
Instead, we'll use a continuous line to represent probability.

* **Continuous uniform distribution**

Flat line since there is same probability of waiting any time (0 – 12 minutes)



* **Still, Probability = area**

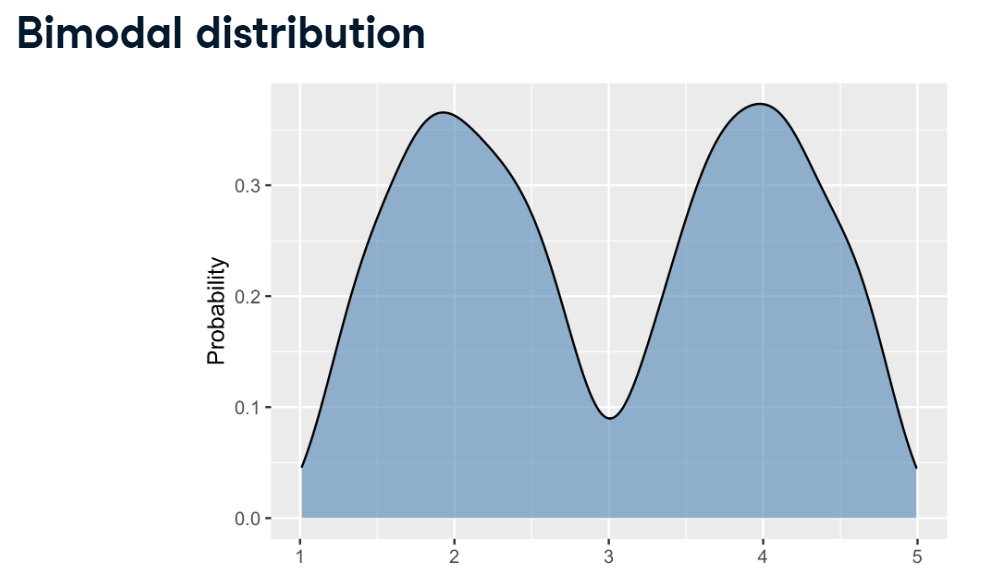


Probability of waiting 12 minutes = 12-0 x 1/12 = 1 = 100%

Probability of waiting 7 minutes or less = 1 – 7/12 = 5/12 = 41.67%

* **Bimodal distribution**

Non – uniform, where some values have a higher probability than others (e.g. 2 modes, book sales depending on book being hardback or paperback



* **Binomial distribution**

Binary outcome, win or lose, success or failure, 1 or 0

Probability distribution of number of successes in a sequence of independent events

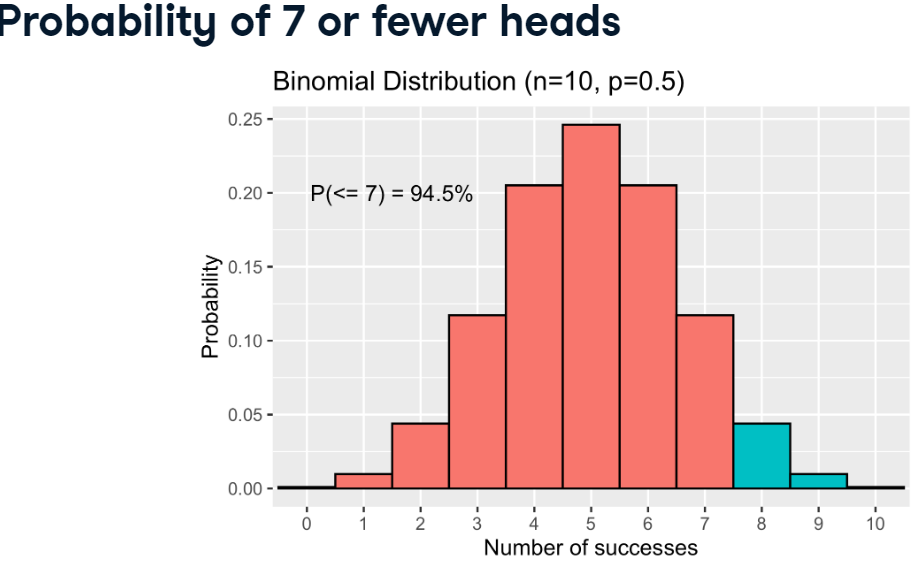
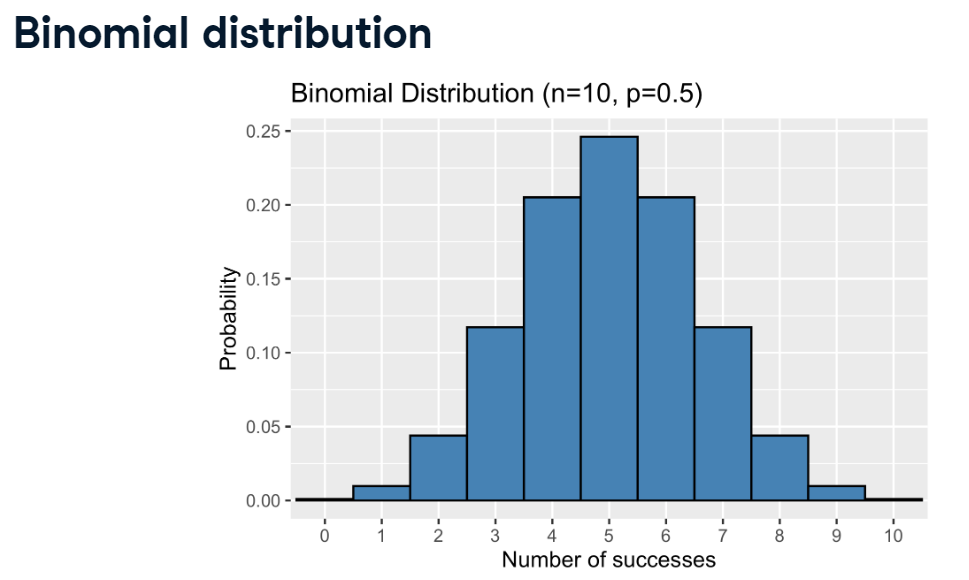
Does not require equal probability for each outcome

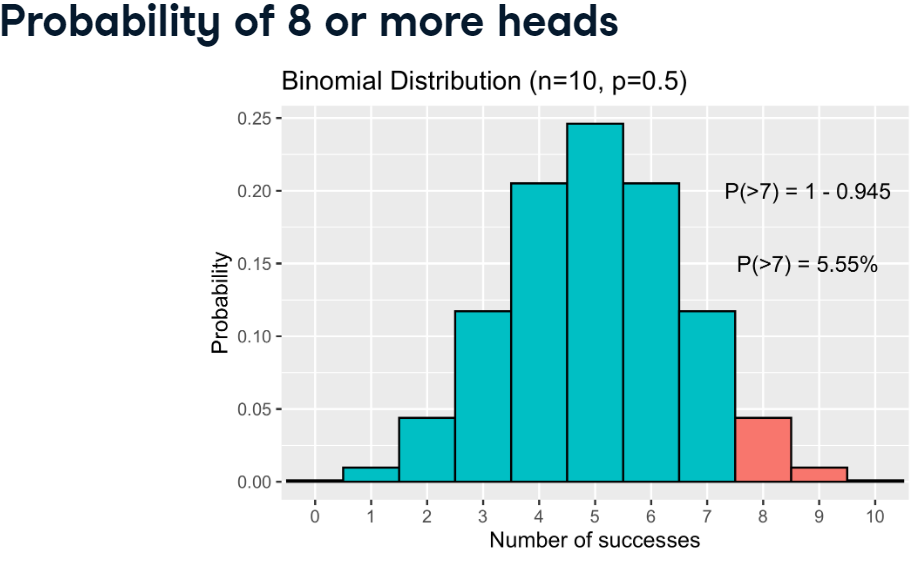
***Described with n (total # of events) & p (probability of success)***

***Expected Value = n x p (bar with highest peak | p = expected value / n***

***Discrete distribution since working with a countable outcome***

**E.g. Probability of getting the number of heads in a sequence of coin flips.**

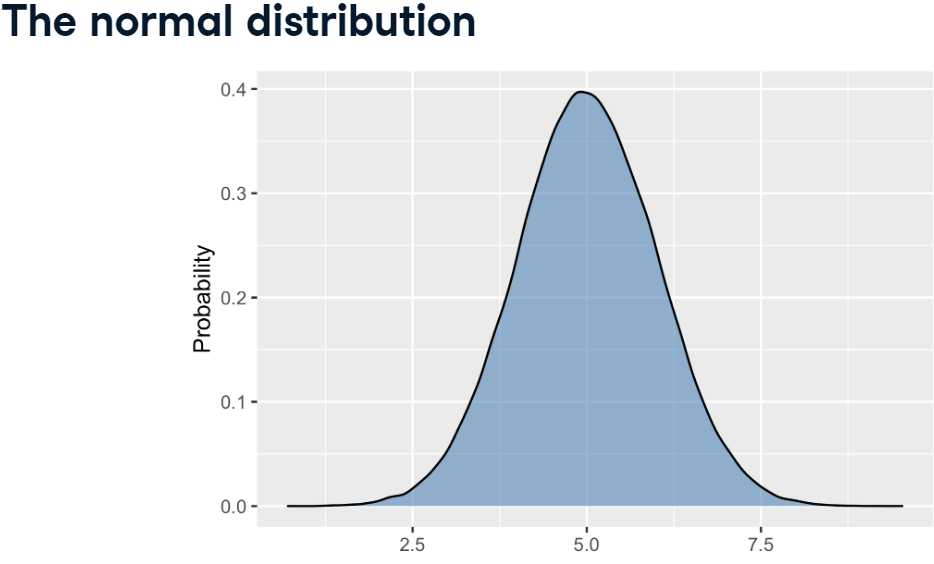




* **Normal distribution - continuous probability distribution**

***Bell shaped curve***

Very common, E.g. blood pressure, retirement age



***Symmetrical (left side is mirror of right side)***

***Total area beneath curve(s) equal 1 as covers all possible outcomes***

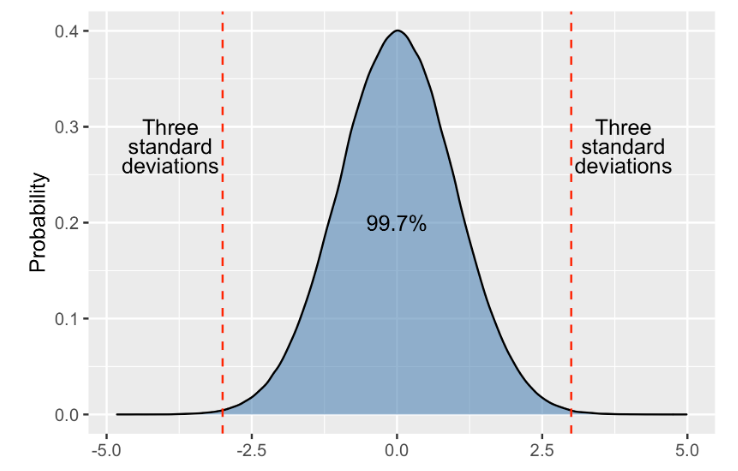
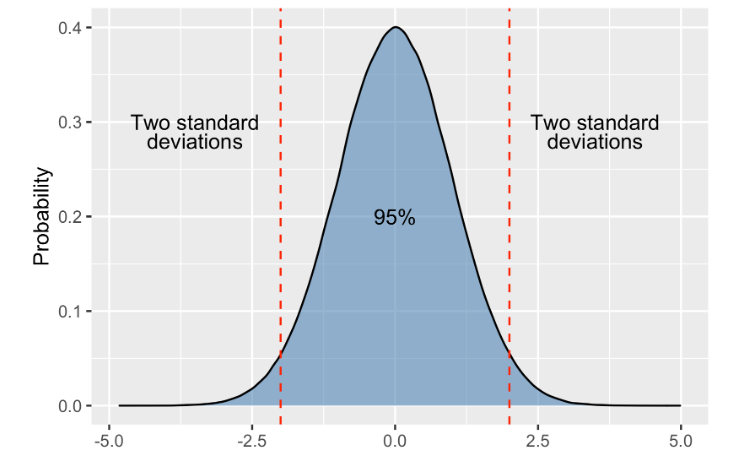
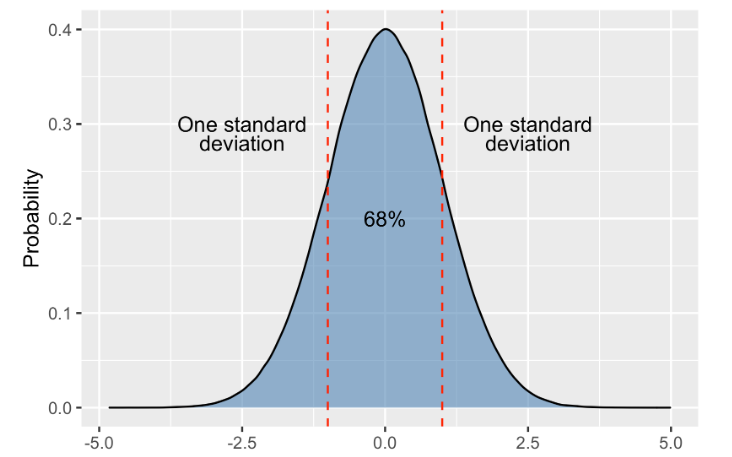
***Probability never hits 0***

* **Areas under normal distribution - 68-95-99.7 rule**

***68.27% of area is within one standard deviation of the mean***

***95.47% of the area falls within two standard deviations of the mean***

***99.7% of the area falls within three standard deviations***

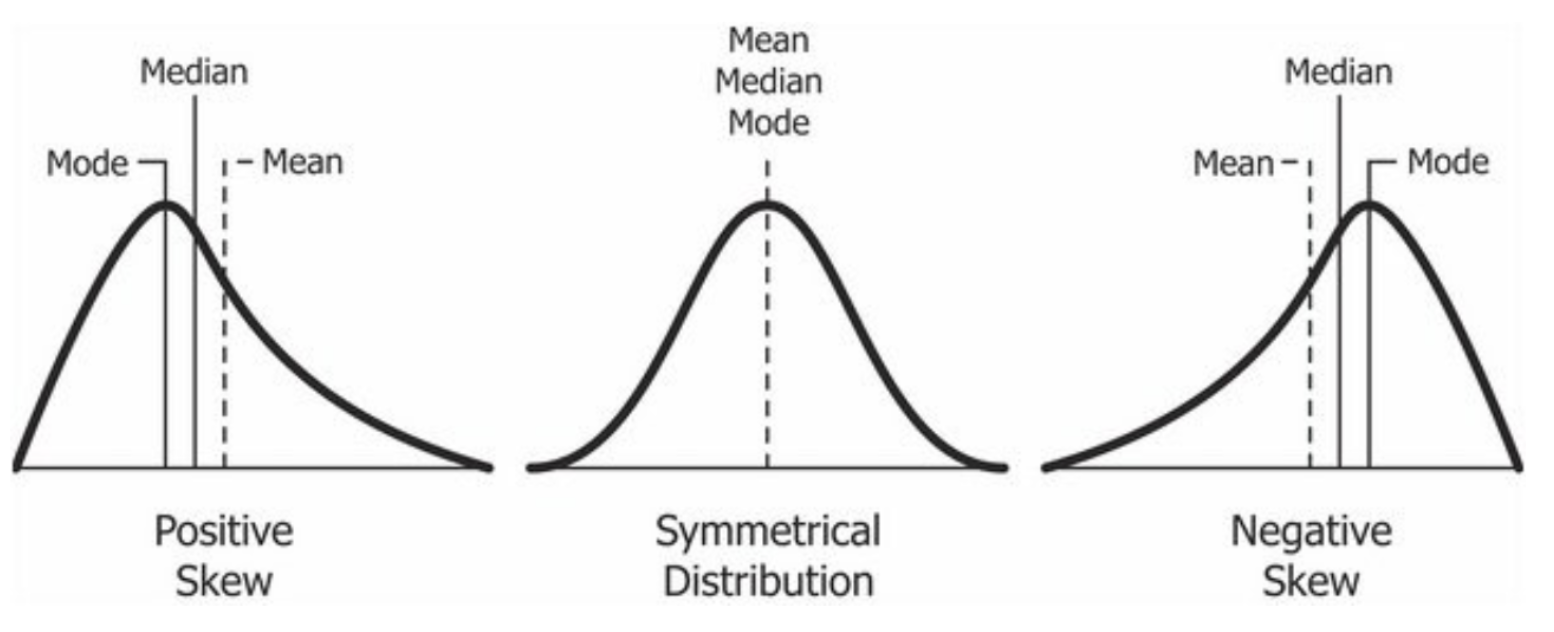


* **SKEW - Skewness - Describes the direction that the data tails off**

***-2 - + 2 is acceptable normal distribution***

***Positively skewed frequency distribution - mean is always greater than median is always greater than mode***

***Negatively skewed frequency distribution - mean is always lesser than median is always lesser than mode***

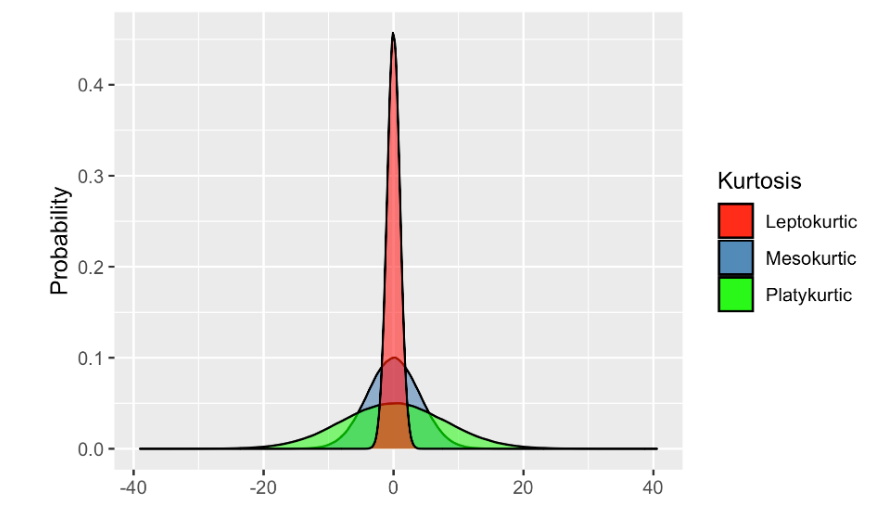
******

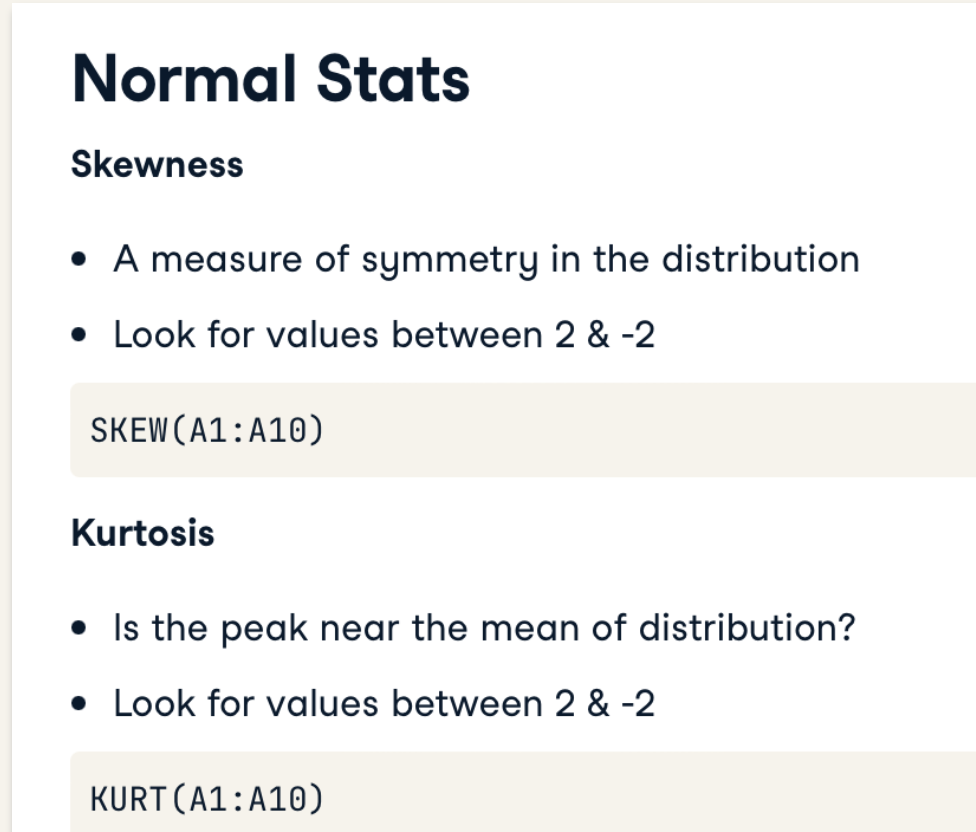
**Right skewed Left skewed**

* **KURT - Kurtosis - Describes the occurrence of extreme values in a distribution**

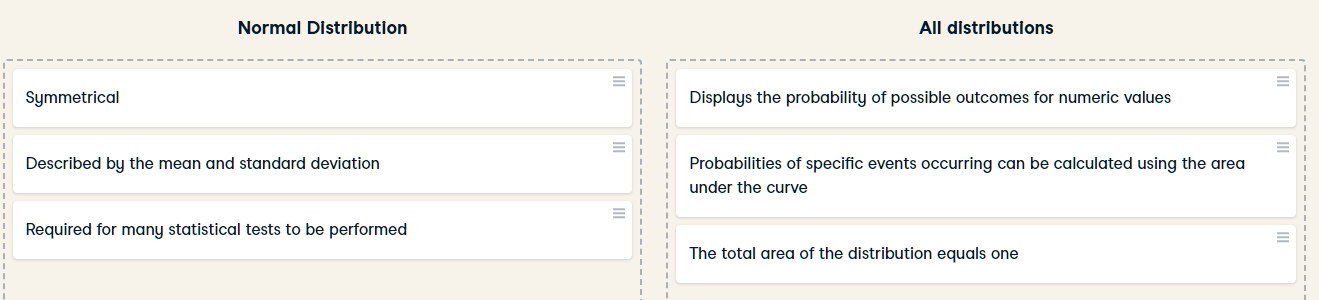
***-2 - + 2 is acceptable normal distribution***

**Positive – Leptokurtic | Normal – Mesokurtic | Negative - Platykurtic**





**Normal distribution vs All distributions**



* **Central Limit Theorem (CLT)**

Sampling distribution will approach normal distribution as sample size increases

Only applies when samples are taken randomly and are independent

Sample size of at least 30 is required for the central limit theorem to apply

Applies to other summary statistics

* **Law of large numbers**

As sample size increases, sample mean will approach theoretical mean (expected value)

Does not always result in sample taking the form of normal distribution as sample size increases

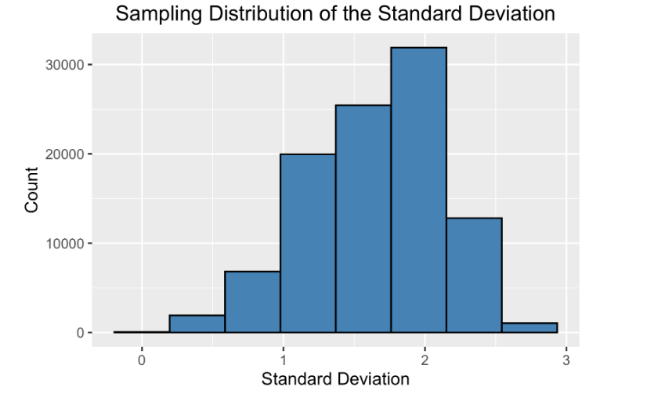
* **Sampling distribution of summary statistics (E.g. Sampling distribution of sample mean)**

Roll die 5 times | Record sample mean | Repeat 10 times

Repeat process for 100, 1000, 10000, 100000, 1000000, 10000000 sample means

***Sampling distribution more closely resembles the normal distribution***

* **Standard deviation and CLT**



* Take STD of 5 rolls 100,000 times
* Sample STD are distributed normally, centred around 1.9 which is the distribution's standard deviation

* **Proportions and CLT**

Sample die 5 times with replacement | Record how many times 4 is rolled | Repeat 1,000 times | Plot distribution of 4s in each sample

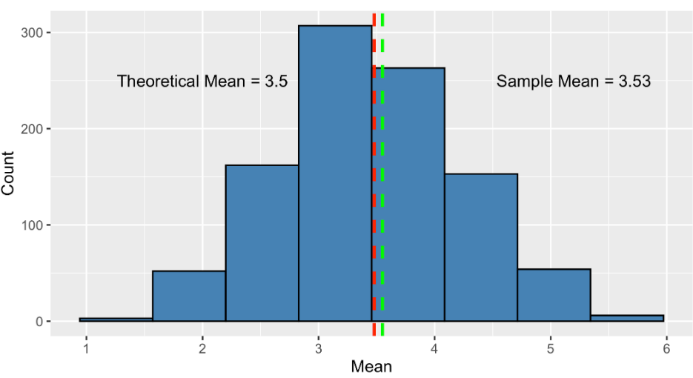
It resembles a normal distribution centred around 0.16, since there is a 1/6 chance of rolling a four

* **Mean of the sampling distribution (example of law of large numbers)**

Since these sampling distributions are normal, their mean can be used to estimate a distribution's mean, standard deviation, or proportion

E.g. Distribution of 1,000 sample means. Red line shows theoretical mean for a dice roll of 3.5 and sample mean of 3.53 as a green dotted line

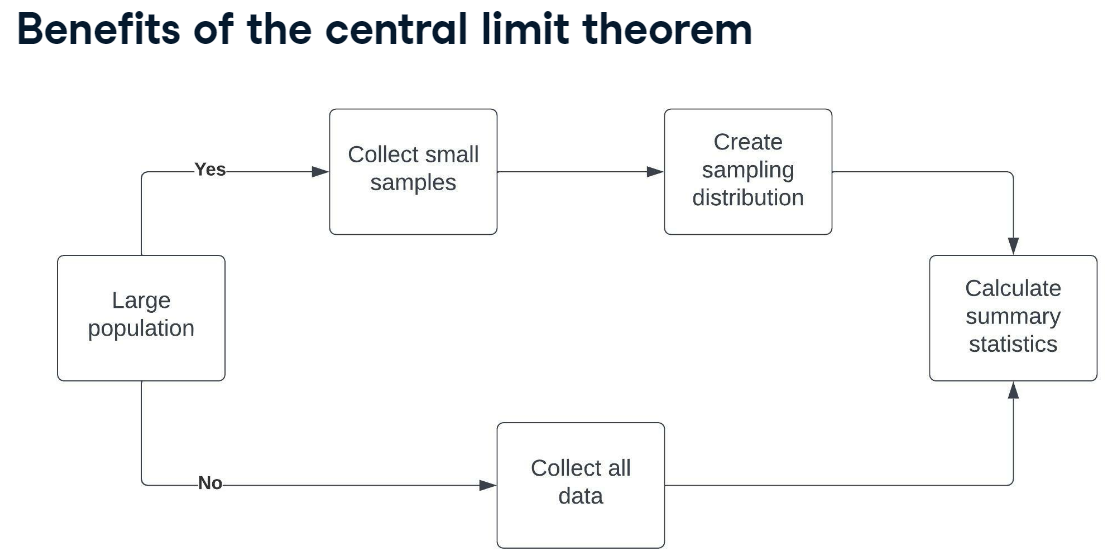
With die rolling examples, we know what the underlying distributions look like for the mean, standard deviation, and proportions, but if we don't, this can be a useful method for estimating characteristics of an underlying distribution



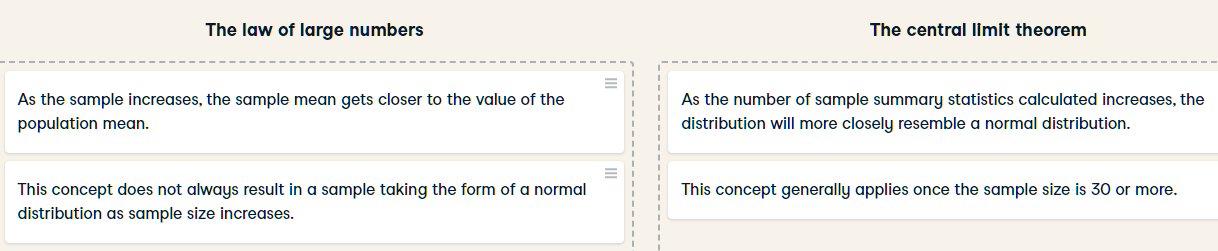
* **Benefits of CLT**

CLT is useful when there is a huge population and no time or resources to collect data on everyone

Instead, collect smaller samples and create a sampling distribution to estimate summary statistics



* **CLT vs Law of large numbers**



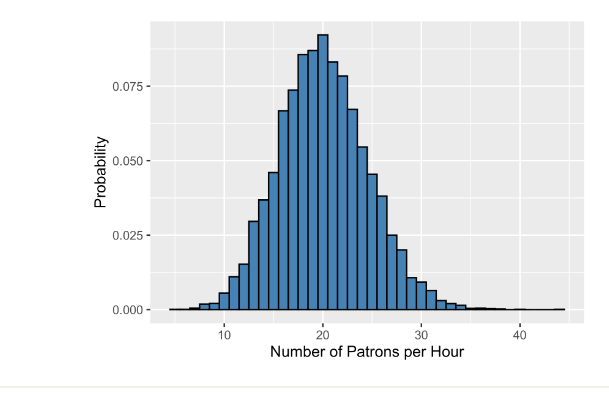
* **Poisson Distribution process - Average # of events in given time period known - time or space between events is random**

# of animals adopted from an animal shelter each week | # of people arriving at a restaurant each hour | # of visits to a website in a day

* **Lambda (λ) = average # of events per time period = probability of some # of events happening over a fixed period of time = expected value of distribution (mean of probability distribution)**

Probability of at least 5 animals getting adopted in a week | probability of 12 people arriving at a restaurant in an hour | probability of > 200 visits to a website in a day

* P**oisson distribution of sample size 200 and λ = average number of patrons per hour = 20**



***This is a discrete distribution - counting events, and 20 is the most likely # of patrons to visit in 1 hour***

* **Lambda is distribution's peak - Lambda changes the shape of distribution**

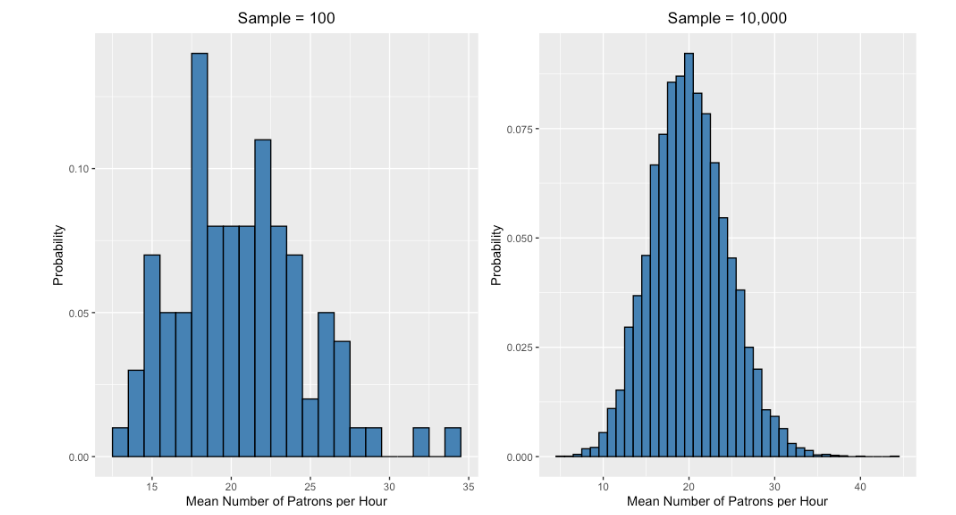
E.g. restaurant patrons per hour - Poisson distribution with **λ** = 10, **λ** = 20 and **λ** = 50



***However, distribution's peak is always at its lambda value***

* **CLT applies to Poisson**

If mean for each is calculated in large # of samples, distribution of sample means of a Poisson distribution looks like the normal distribution

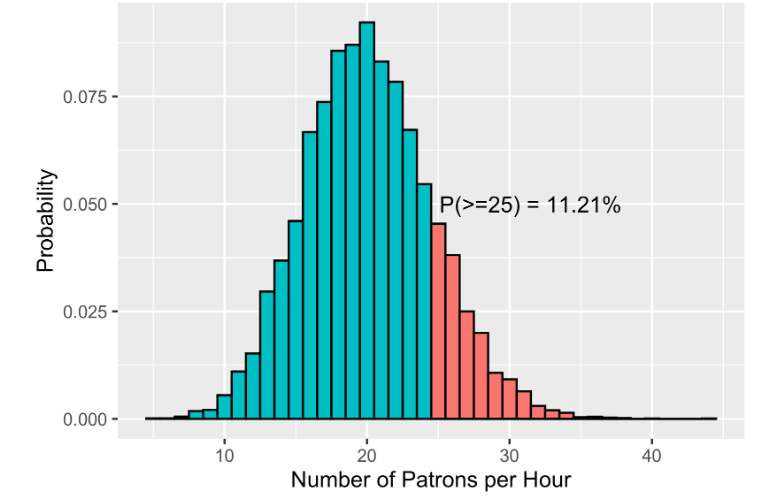
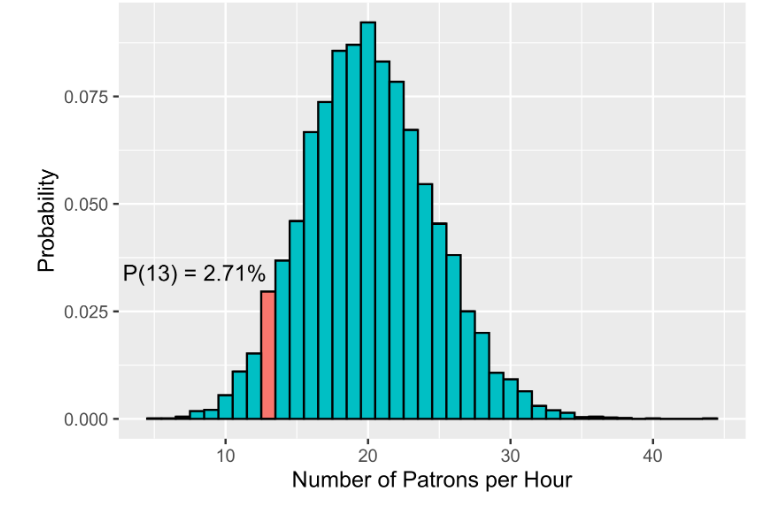


* **E.g. Probability of # of patrons in an hour**

Calculate probability of specific # of patrons visiting restaurant by measuring height of bar representing that value

Probability of exactly 13 patrons visiting restaurant in an hour, given the average is 20 patrons, is 2.71%

Probability of at least 25 patrons visiting in an hour = add heights of all bars from 25 onwards, is 11%



* **Hypothesis testing - Group of theories, methods, and techniques to compare populations**

History of hypothesis testing is well-established

Early origins in 18th century when analysis of birth records showed that each birth has slightly larger probability of being male than female

E.g. Theory that increasing price of product will increase revenue

E.g. Changing name of website will increase traffic

E.g. Use hypothesis testing to analyse whether a medication is effective in treatment of specific health conditions

* **Null hypothesis and Alternative hypothesis**

1. Null Hypothesis - H0 - Status Quo

Start with assumption that no difference exists between populations to reduce risk of bias

E.g. Null hypothesis = there is no difference in gender birth ratio between women who do and do not take vitamin C

1. Alternative hypothesis - H1 - Challenger statement

Can take 1 of 2 forms

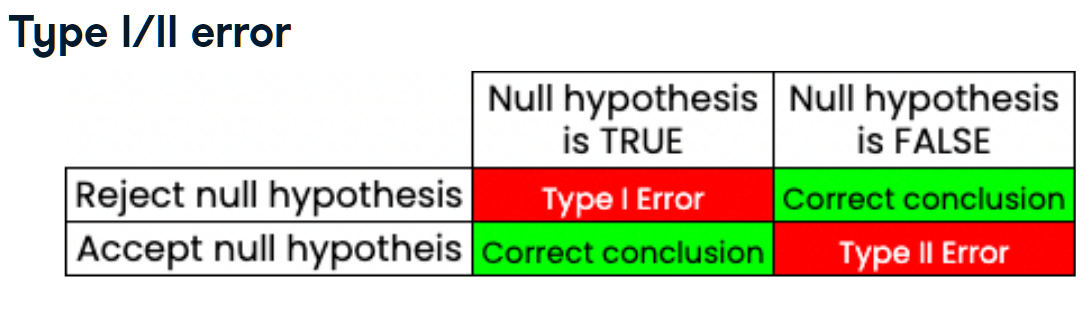
* 1. Sate there is difference between male and female births in women taking vitamin C vs those who do not

OR

* 1. State direction of difference - population taking vitamin C have more female births vs those who do not

* **Hypothesis testing workflow**

1. Define target population to analyse the difference between - adult women using or not using vitamin C
2. Define null hypothesis - births are equally likely to be male or female in both populations
3. Define alternative hypothesis - babies are more likely to be female in women taking vitamin C
4. Collect sample data - gender status of babies born in both populations
5. Perform statistical tests on sample data
6. Draw conclusions about the population that the sample represents
7. Reject Null Hypothesis OR Fail to Reject Null Hypothesis



* **How much data is needed**

Applying CLT, with larger sample size the mean number of male and female births approaches the population means

However, collecting large samples can take a lot of time and resources

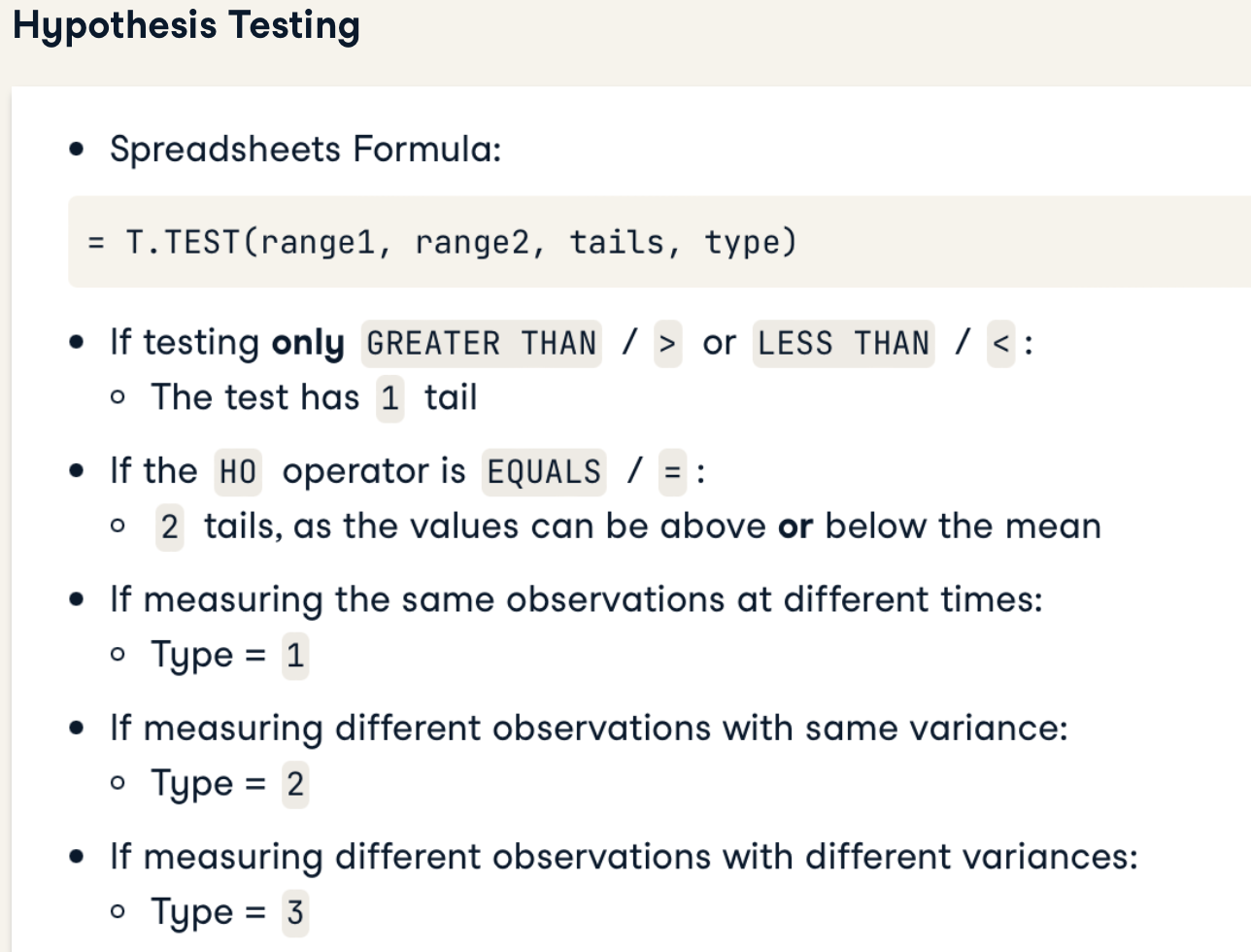
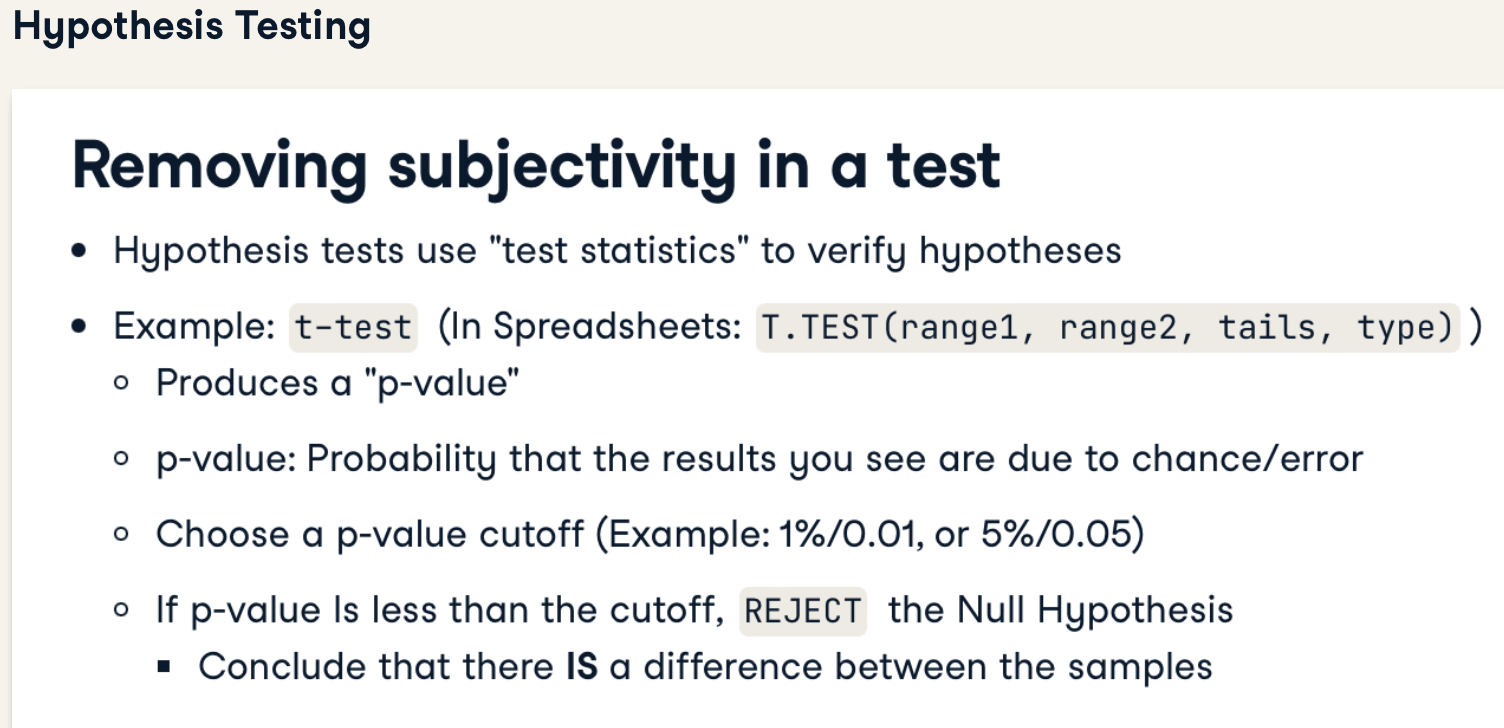
Common approach is to look at peer-reviewed research on similar hypothesis tests to find out how large the samples were – this can then serve as a benchmark

* **T - TEST Hypothesis Testing in Spreadsheets**

***T-Test to make sure sample is same or different to population or other samples***

***Infers if there is a difference between 2 means***

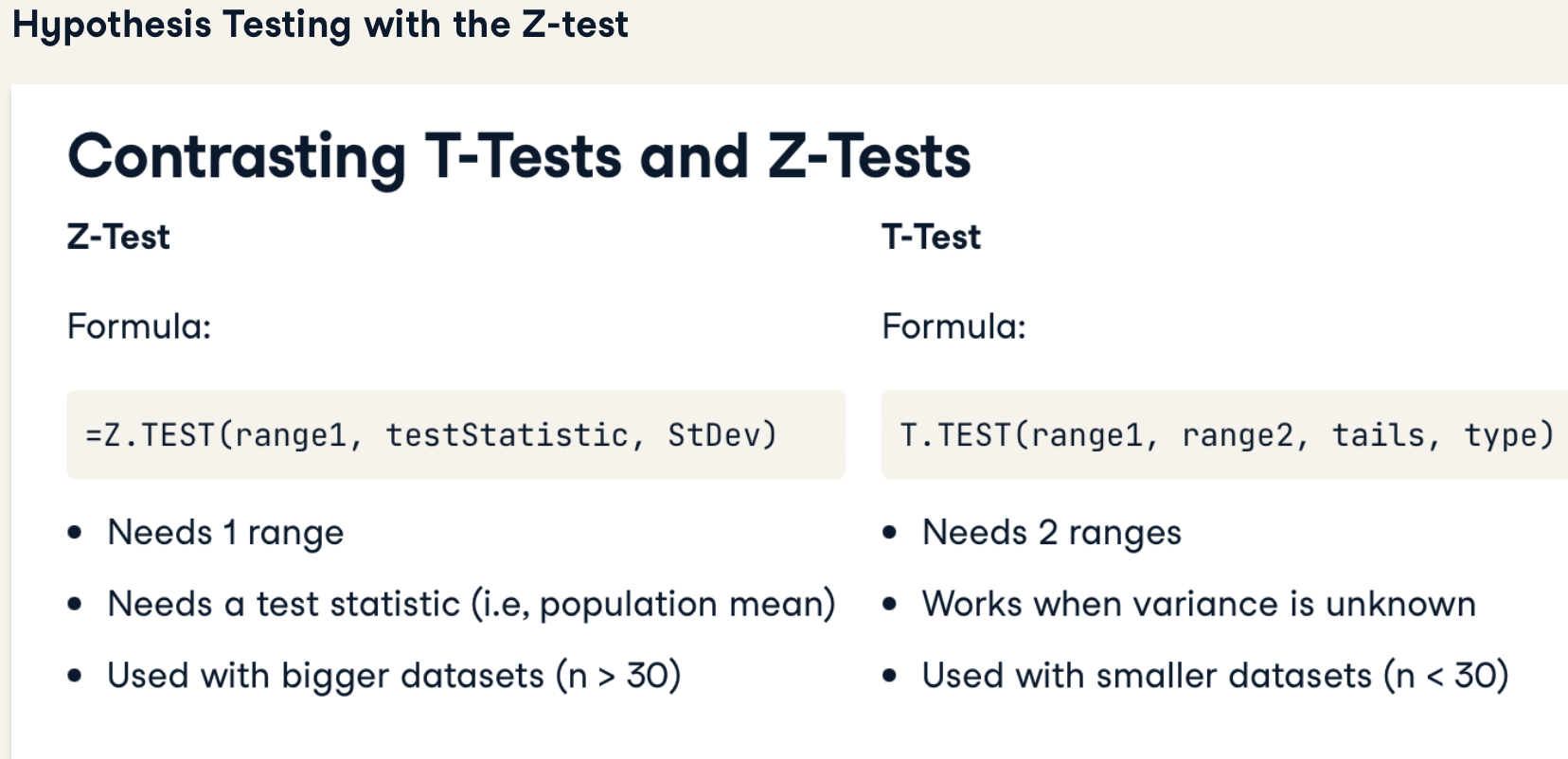
1. Determine whether 2 population means are statistically different
2. Select p-value cutoff prior to test (usually 0.05)
3. If p-value < 0.05 - Reject H0
4. If p-value>= 0.05 - Fail to Reject Ho

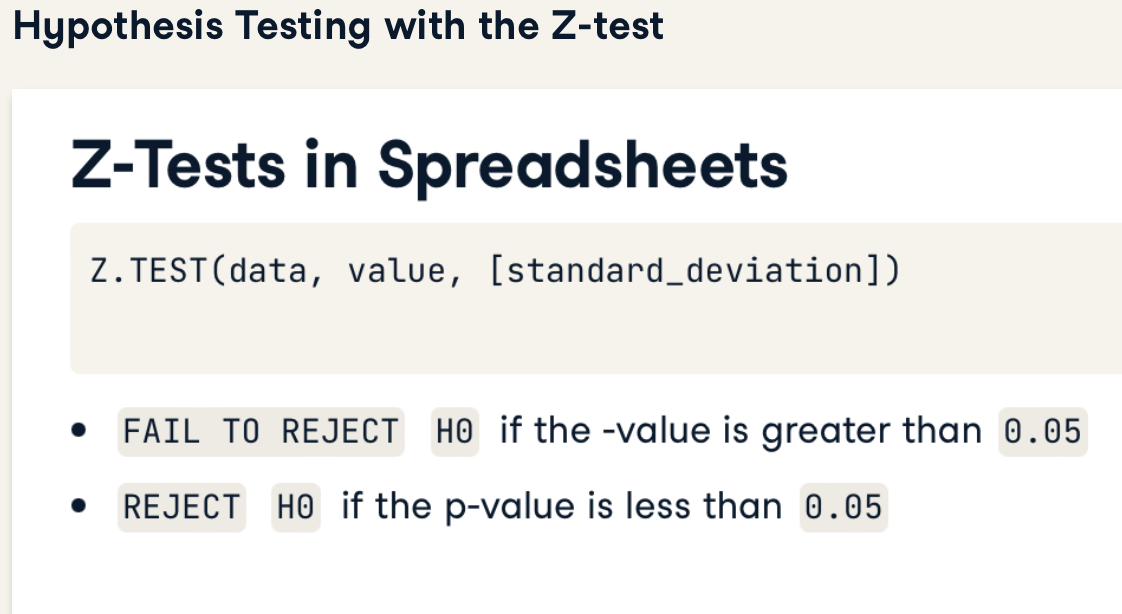


* **Z - TEST Hypothesis Testing in Spreadsheets**

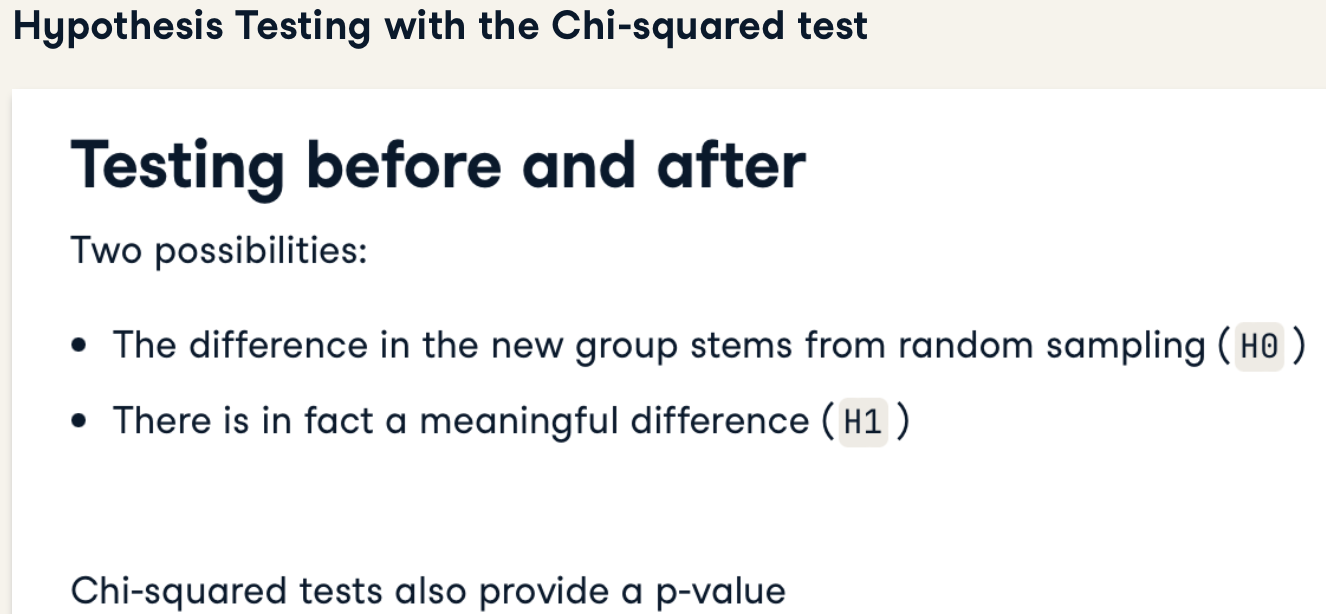
***Z - Tests calculates the probability of value provided***

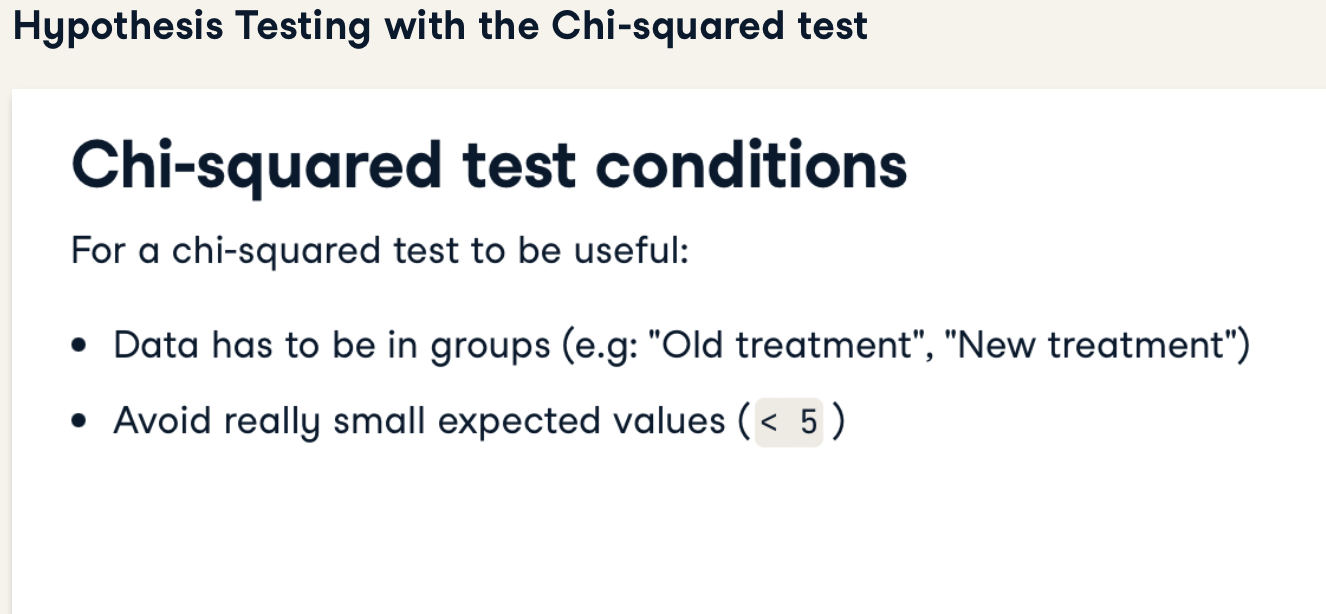
***Multiply by 2 if two-tailed (above or below average)***

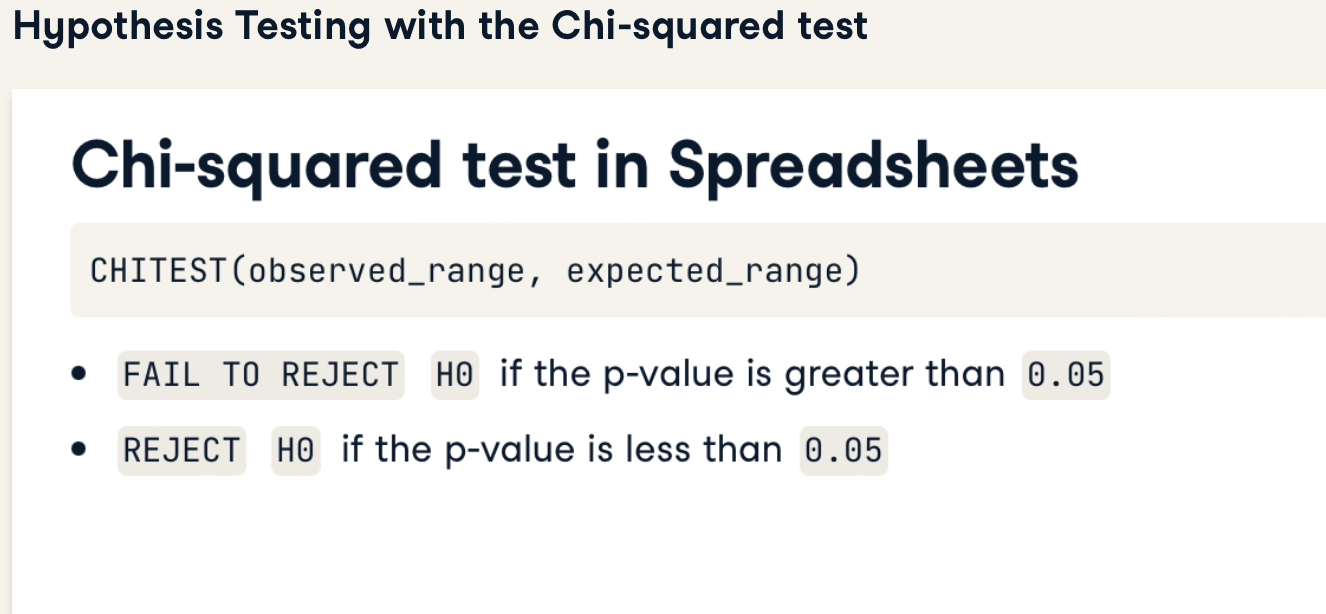
******

******

* **Chi - Squared Hypothesis Testing in Spreadsheets**

****

****

****

* **Independent and dependent variables terminology in hypothesis testing**

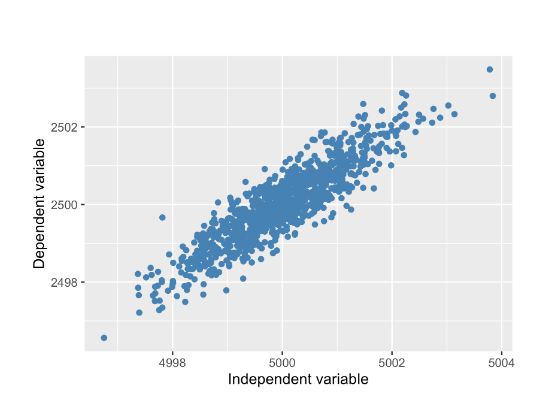
Data is defined in terms of difference expected to be observed in the alternative hypothesis

**Independent variable -** data expected not to be affected by other data

Vitamin C in gender birth ratio hypothesis test - meaning it is independent of male to female birth ratio

**Dependent variable** - data expected to be affected by other data

Birth gender in alternative hypothesis - proposes birth gender ratio will be affected by vitamin C (dependent on vitamin C)



***independent variable is always on the x-axis and the dependent variable is on the y-axis***

* **Hypothesis testing experiments - Subset of hypothesis testing**

Aims to answer a question E.g. What is the effect of the treatment on the response?

Treatment = Independent variable

Response = Dependent variable

**E.g. Advertising as a treatment**

Question – what is effect of advertisement on # of products purchased

Treatment (Independent variable) = advertisement

Response (Dependent variable) = # of purchased products

* **Controlled experiments**

Participants are randomly assigned to either treatment group or control group

Treatment group will see advertisement

Control group will not see advertisement

Groups must be comparable to avoid incorrect conclusions

If groups not comparable (e.g. different ages), will influence results - experiment biassed

* **Gold standard of experiments - Ideal experiment will eliminate as much bias as possible**

**Principle: the fewer the opportunities for bias in experiments, more reliably conclude whether the treatment affects the response**

**Randomised controlled trial / Randomization experiment protocol** - participants are randomly assigned to treatment or control group to ensure groups are comparable

**Blind trial / Blinding** - participants don't know if they're in the treatment or control group to ensure the effect of treatment is due to treatment and not the idea of getting treatment E.g. using a placebo – common in clinical trials

**Double-blind randomised controlled trial -** person administering treatment / experiment also unaware whether administering treatment or placebo to protect against bias in response as well as analysis of results

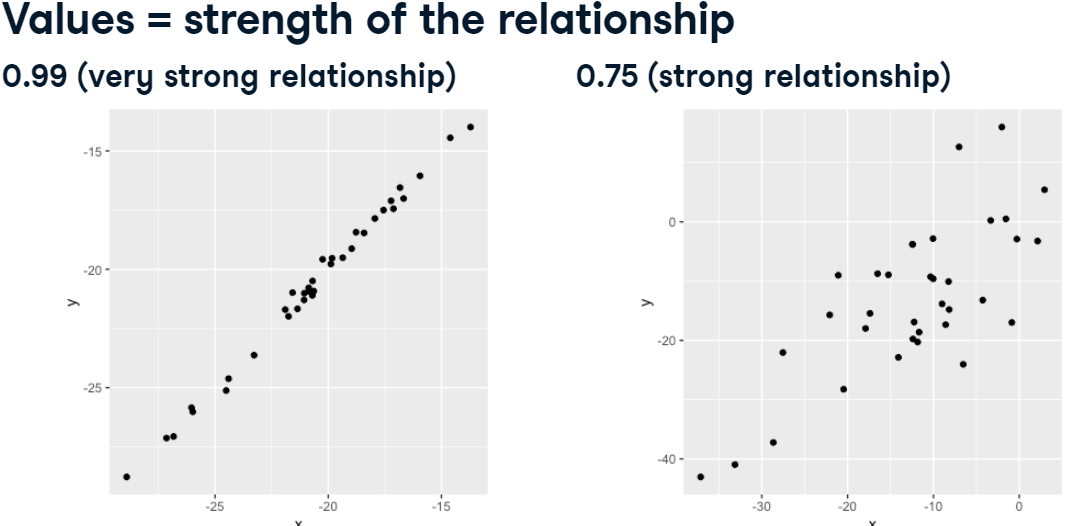
* **Randomised Controlled Trials vs. A/B testing**

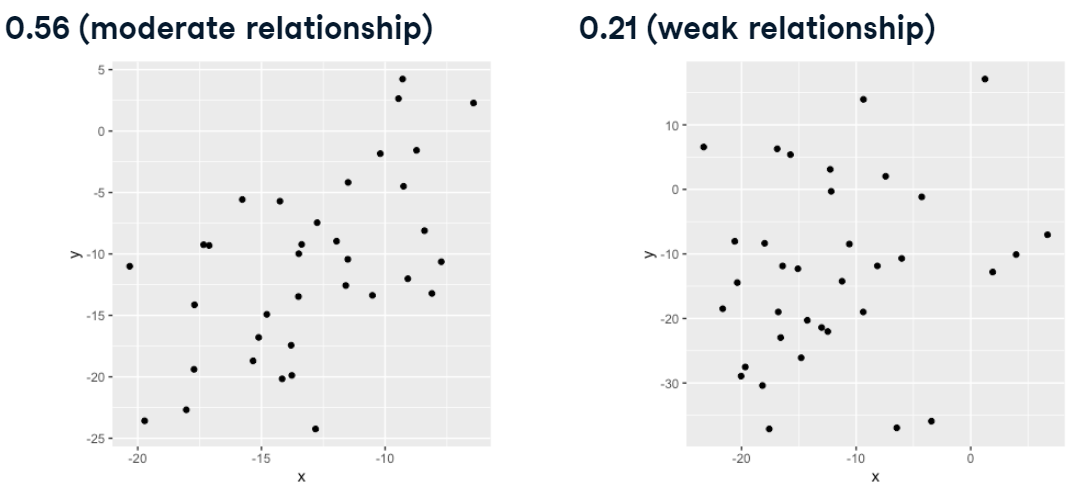
Randomised controlled trials can have multiple treatment groups - academia, scientific and clinical research

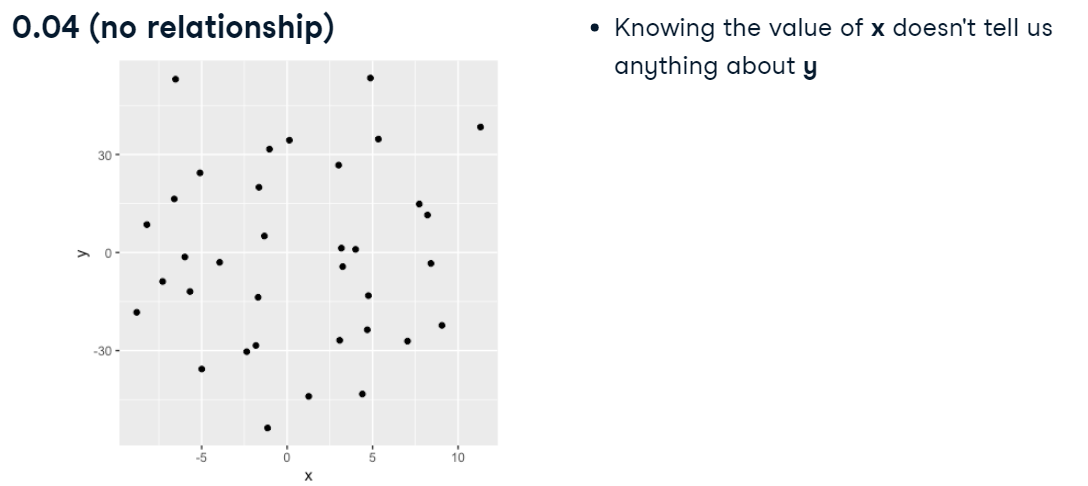
* Referred to as A/B testing when only splits population into two groups of treatment and control - used in industries
* **Correlation (Pearson Correlation Coefficient) - Quantifies strength of relationship between 2 variables**

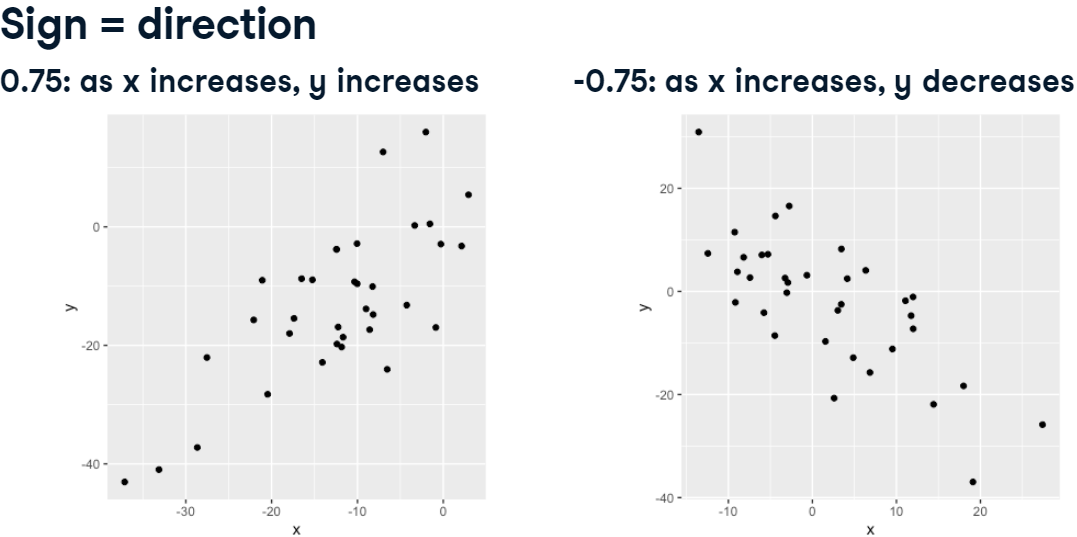
Between –1 and +1 - Magnitude corresponds to strength and (+ or -) corresponds to direction of relationship

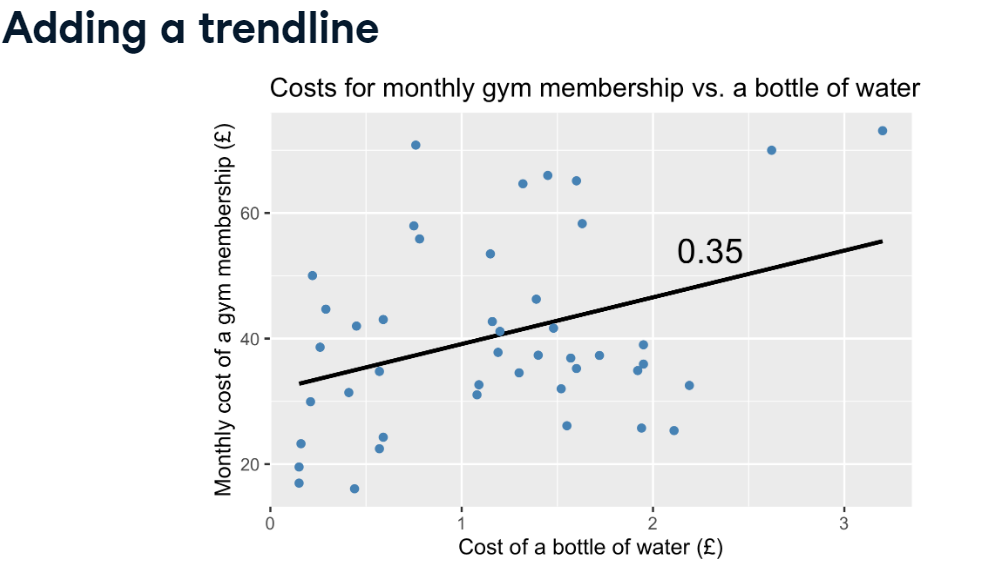
***Only for linear relationships (proportionate changes between dependent and independent variables)***











* **Correlation does not equal causation**

Correlation coefficient 0.61 suggesting moderate positive relationship in plot of life expectancy and cost of water bottle

Does not mean increasing cost of water will increase life expectancy

***Important to distinguish that just because a relationship exists, it doesn't mean that changes in one will result in a in the other***

* **Confounding variables - something that affects the data, but was not accounted for when assessing the relationship between variables**

Ask what else could be affecting values when looking at relationships among data

Cost of water bottle typically higher in stronger economies which may offer high quality healthcare access - So perhaps life expectancy is not affected by cost water but actually affected by the strength of the economy

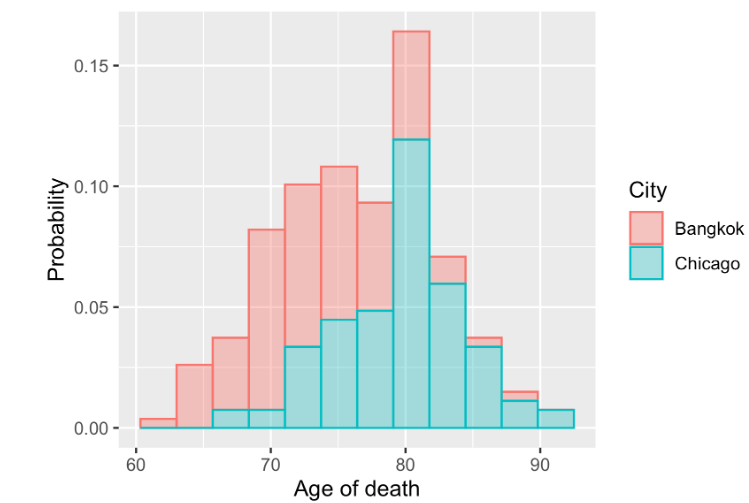
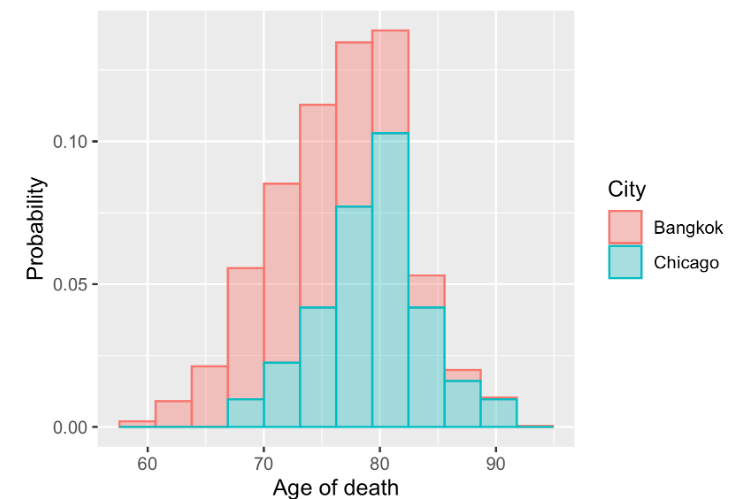
* **Interpreting hypothesis test results - Life expectancy in Chicago vs. Bangkok**
* Test if there is a difference in life expectancy in Chicago and Bangkok
* Null hypothesis - no difference exists
* Alternative hypothesis - Chicago residents have longer life expectancy

1. **Sampling distribution**

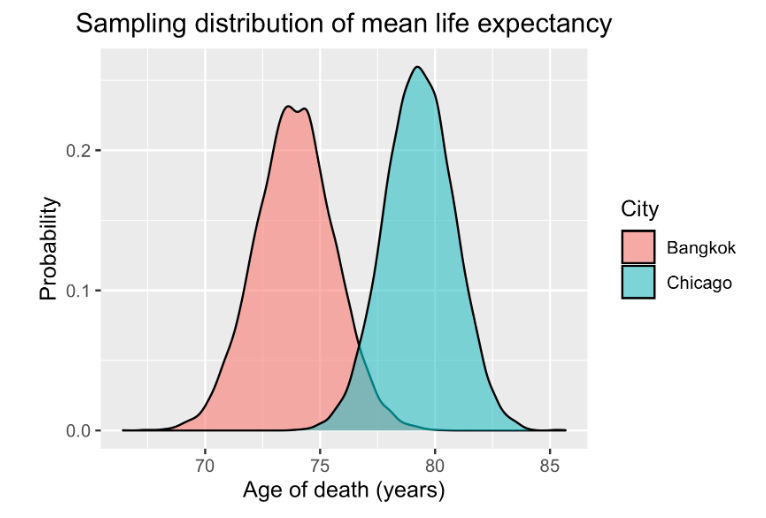
* Collect data on the age of death from 100 residents each in Chicago and Bangkok
* Histogram shows life expectancy sample distributions for each city
* How to know these are the actual mean values for each population?

1. **Different samples**

* Collect age of death data from 100 more residents in each city
* This time there are different results
* How to be sure the difference in life expectancy exists, or are results due to chance? - **Put another way, do the samples truly represent these populations?**

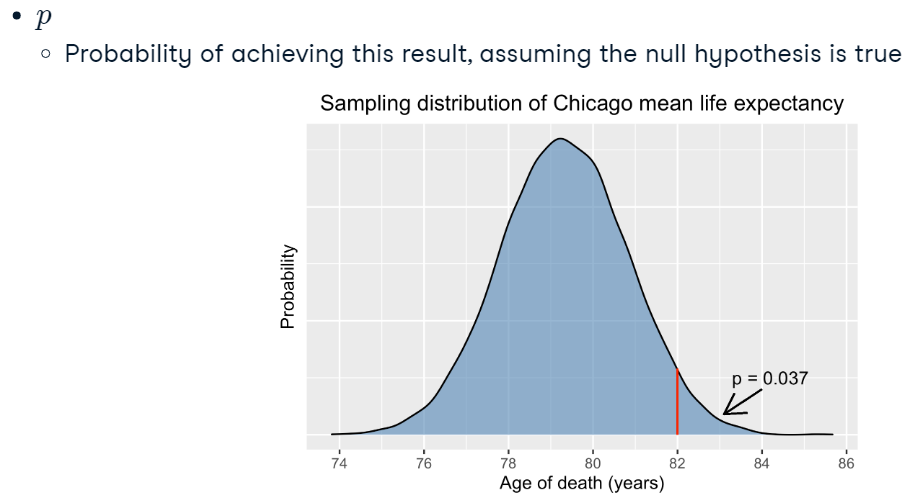


* **Sampling distribution of mean life expectancy**
* One approach is to perform sampling with replacement on original data from each city and calculate the mean life expectancy for each sample as nt possible to collect entire population data
* Repeating this 10,000 times and visualising the results - normal distributions for mean life expectancy in Bangkok and Chicago, and Chicago has larger expected value
* So, can it be concluded that a difference in life expectancy truly exists?

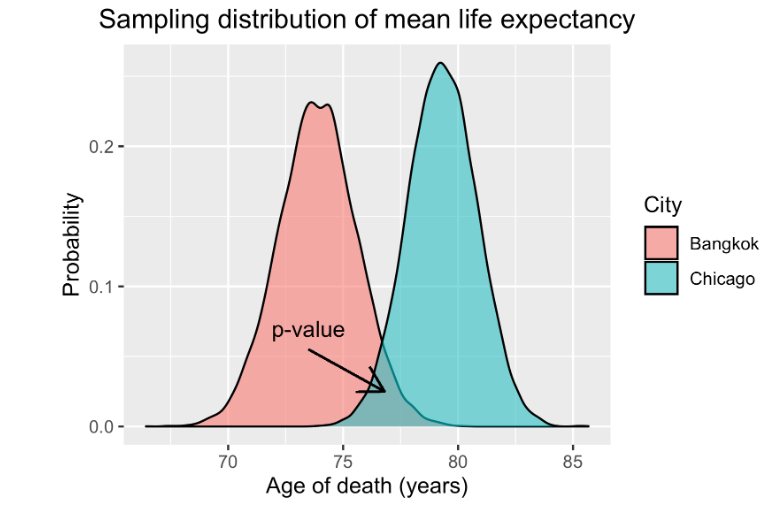


**p-value**

* When drawing conclusions in hypothesis testing we use a metric called a p-value.
* This is the probability of achieving a result at least as extreme as the one we have observed, assuming the null hypothesis is true.
* Suppose we want to know the probability of a sample mean for Chicago life expectancy being more than or equal to 82, given a population mean of 79.3.
* We can visualise the sample means distribution and look at the total area from 82 onwards to determine the p-value of 0.037, meaning there is a 3.7% chance of observing a mean life expectancy of 82 or more.



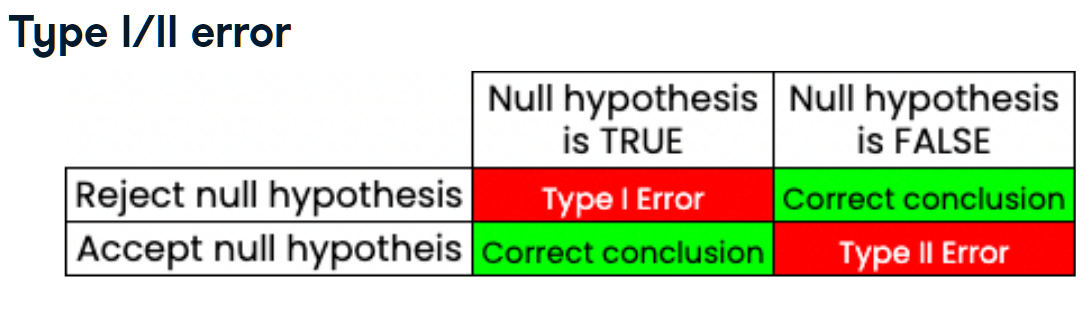
* **p-value**
* Visualise p-value for 2 sample mean distributions = total area that overlaps between them
* How small an overlap is needed to be confident in conclusion?



* **Significance level (α) = Probability threshold / Significance level**
* To reduce risk of drawing a false conclusion - set a probability threshold for falsely rejecting null hypothesis
* Decided before collecting data to minimise bias

Typical value is 0.05 - there is 5% chance of wrongly concluding that Chicago residents live longer than Bangkok residents

* After data collection, look at whether p-value is less than or equal to α
* If p-value meets this criterion - confident in rejecting null hypothesis
* **If p <= α, reject the null hypothesis - describe results as being statistically significant**
* **Type I/II error In hypothesis testing - 4 potential conclusions based on null hypothesis**
* Type I error. - wrongly reject the null hypothesis when it is actually true
* Type II error - wrongly accept the null hypothesis when it is actually false
* Correctly accept the null hypothesis when it's true
* Correctly reject the null hypothesis when it's false



* **Drawing a conclusion**
* Having set alpha for 0.05, now draw a conclusion based on our sample mean distributions
* Overlap of distributions is less than threshold for alpha - meaning likelihood of difference in mean life expectancy between the two cities occurring by chance is less than 5%
* **Therefore, reject null hypothesis and reasonably conclude that mean life expectancy in Chicago is higher than in Bangkok**

