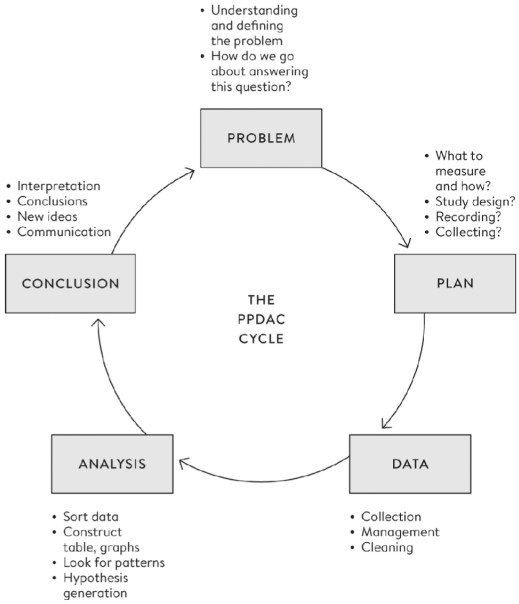
GEA1000 Quantitative Reasoning with Data

▼ PPDAC Cycle

The figure below is a representation of the data problem-solving cycle, “Problem, Plan, Data, Analysis and Conclusion”.

Features of PPDAC

- (to) document the stages a person would undertake when solving a problem using numerical evidence,
- using data which they had collected themselves, or from existing (public) data sets,
- (where) analysis methods can include machine learning algorithms, as well as more traditional statistical techniques.



▼ Chapter 1

Population: The population is the entire group (of individuals or objects) that we want to know something about.

Research Question: The research question is usually one that seeks to investigate some characteristic of a population.

Type of Research Questions	Examples
Make an estimate about the population	What is the average number of hours that students study each week?
	What proportion of all Singapore students is enrolled in a university?
Test a claim about the population	Is the average course load for a university student greater than 20 units?
	Does the majority of students qualify for student loans?
Compare two sub-populations	In university X, do female students have a higher GPA score than male students?
	Are student athletes more likely than non-athletes to do final year projects?
Investigate a relationship between two variables in the population	Is there a relationship between the average number of hours students spend each week on Facebook and their GPA?
	Does drinking coffee help students pass the math exam?

▼ Sampling

Exploratory Data Analysis(EDA)

1. Generate questions about our data
2. Search for answers by visualising, transforming, and modelling data
3. Use what is derived from the data to either:
    a. Refine our existing question
    b. Generate new questions

Population vs Sample

Population of Interest – A group in which researcher has interest in drawing conclusions of the study.

Population Parameter – a numerical fact about a population.

Sample – A proportion of the population selected in the study.

Estimate – An inference about the population's parameter, based on information obtained from a sample.

Sampling Frame

Sampling frame is the "Source Material" from which sample is drawn. That is, sample is drawn from the sampling frame.

It may not cover the population of interest, or may contain some units which are not in the population of interest.

One of the essential conditions for generalisability is that the sampling frame should be greater than or equal to the population of interest. In other words, the intersection of the population of interest and the sampling frame should be the population of interest

Census vs Sample

Census – An attempt to reach out to the entire population of interest

Sample – A proportion of the population selected

Why sampling over population data?

1. Cost
2. Speed

Bias

Selection Bias	Non-response Bias
Associated with the researcher's biased selection of units	Associated with the participants' non-disclosure of information related to the study
• Imperfect Sampling Frame • Non-Probability Sampling	• Disinterested • Inconvenient • Unwilling to disclose sensitive information

Probability Sampling

The key point to remember is that in probability sampling, when you repeat the process of sampling, it is possible to obtain a different sample. That is, the choice of sample relies on probability and chance - it is not deterministic. You cannot predict (with high accuracy) the sample that will be chosen in a probability sampling method before the process begins.

So, the selection process is via a **known** randomised algorithm. (But knowing the randomised algorithm doesn't make it any less random)

The probability of selection may not be the same throughout all units of the population

Probability sampling eliminates selection bias but there is no guarantee of non-response bias being eliminated (In other words, probability sampling and non-response bias are unrelated and independent - since non-response bias depends on the participants chosen, not how the participants were chosen)

There are 4 main ways to conduct probability sampling:

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling

Simple random sampling (SRS)

- Units are selected randomly **without replacement** from the sampling frame
- **Mechanism**: Random Number Generator (Example: Random-Digit Dialling)
- An SRS of size n consists of n units from the population chosen in such a way that every set of units has equal chance to be the sample actually selected.
- Sample results do not change haphazardly from sample to sample (because you have eliminated a lot of the confounders by making the process random - so, on average, all the features of units are random too). Any variability in the results is purely due to chance.
- Each unit has an equally likely and known chance of being chosen. That is, you can calculate the probability of a unit being chosen before the sampling process begins.

Advantage: Sample tends to be a good representation of the population

Disadvantage: Subject to non-response; accessibility of information

Systematic Sampling

A method of selecting units from a list by applying a selection interval *K, and random starting point from the first interval*.

That is, you choose a fixed interval K and a random point in the among the first K units (say, p). Then, your sample consists of $p, p+K, p+2K, \ldots$

In other words, you have exactly 1 unit in the sample from each interval of length K. This ensures that no interval is left behind unrepresented in the sample.

An example would be as follows: You are trying to collect data from houses. You pick a random house to start with. Then, you go to every 3rd house starting with that house.

The reason why this is random is because different random starting points can be produced, which give rise to different samples.

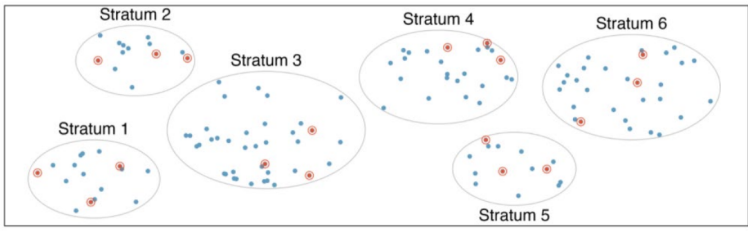
Advantage: Simpler selection process than simple random sampling

Disadvantage: May not be representative of population if list is non-random. (imagine if your list of houses is arranged in such a way that every 3rd house belongs to a poor family. Then, if you choose $K = 3$ and start with a house that belongs to a poor family, your entire sample will just consist of poor households - not representative of everyone else)

Stratified Sampling

- Break down the population into strata.
- Each stratum are similar in nature but size may vary across strata

- Then, apply simple random sampling from every strata
- Examlple: Sample count in a general election

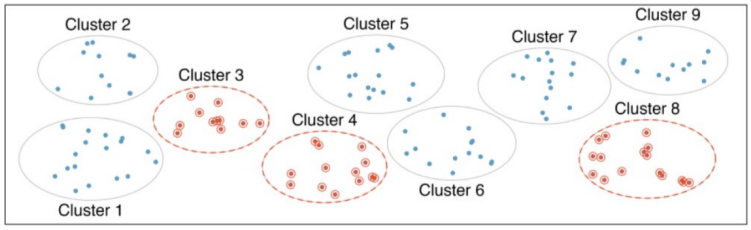


Advantage: able to get a representative sample from every stratum

Disadvantage: Need information about sampling frame and stratum (How to split the sampling frame into strata is an important question for any stratified sampling procedure)

Cluster Sampling

- Break down the population into clusters
- Randomly sample a fixed number of clusters
- Include all observations from selected clusters
- Example: Mental wellness surveys in school



Advantage: Less time-consuming, less tedious, and less expensive

Disadvantage: High variability due to dissimilar clusters or small number of clusters

Summary of Probability Sampling Methods

Sampling Plan	Advantages	Disadvantages
Simple Random Sample	Good Representation of the Population	Time-consuming; accessibility of information
Systematic Sample	Simpler selection process as opposed to Simple Random Sampling	Potentially under-representing the population
Stratified Random Sample	Good Representation of Sample by Stratum	Require Sampling Frame and criteria for classification of population into stratum
Cluster Random Sample	Less time-consuming and less costly	Require larger sample size in order to achieve low margin of error

Non-probability sampling

When the selection of individuals/unit were not done by randomisation, but by human discretion, we call it non-probability sampling.

Example 1: **Convenience Sampling** - It is a non-probability sampling method in which the researcher uses the subjects that are most easily available to participate in the research study.

- An example of convenience sampling: Mall surveys.
- Issue 1: Demographics of mall goers – teenagers, retired people, people who are more affluent. Other groups (non-teenagers and retirees, and the
  not so affluent) are left out. This is a good example of **selection bias**.
- Issue 2: Individuals asked to do the survey may not respond. This could lead to **non-response bias.**

Example 2: **Volunteer Sampling** - It is a non-probability sampling method in which the researcher *actively* seeks volunteers to participate in the study.

General steps involved in Sampling

1. Choose Sampling frame
2. Sample from sampling frame
3. Remove unwanted units (those that do not fall under the population of interest)

Generalisibility Criteria

It is important to know when you can (and more importantly, when you cannot) generalise the results or findings using your sample to the population of interest. Only when all 4 of the following criteria are met, you can safely generalise the claims to a larger population.

1. **Good sampling frame**. (the sampling frame must be at least as large as the population of interest)
2. **Probability based sampling** (you cannot use convenience sampling or volunteer sampling. Your sample must be randomly chosen from the sampling frame)
3. **Large sample size** (if you have data from a sample size of 50, you probably cannot generalise it to a population whose size is 1000)
4. **Minimum non-response** (the more the non-response, the less accurate/complete data you will have)

▼ Variables

A **variable** is an attribute that can be measured or labelled.

In research questions involving examining relationships between variables there are typically 2 sets of variables, namely **independent variables** and **dependent variables**.

An **independent variable** is a variable that maybe subject to manipulation (either deliberately or spontaneously) in a study.

A **dependent variable** is a variable which is *hypothesised to change* depending on how the independent variable is manipulated in a study.

Research question	Dependent variable/Independent variable
Do NUS students who make notes using pen and paper score better in GEA1000 than those who use laptops?	Independent variable : Method of note taking for GEA1000
	Dependent variable : GEA1000 grade.
Does amount of caffeine consumed per day affect the quality of sleep amongst Singaporean adults?	Independent variable : Amount of caffeine consumed per day
	Dependent variable : Quality of sleep

There are two main types of variables: **numerical** and **categorical**.

Categorical variables take category or label values. Each observation can be placed in only one label, and the labels are mutually exclusive (i.e no 2 labels overlap with each other). Example : Smoking status can be a categorical variable, with two groups (smoker or non-smoker). Education Level is another example with multiple labels.

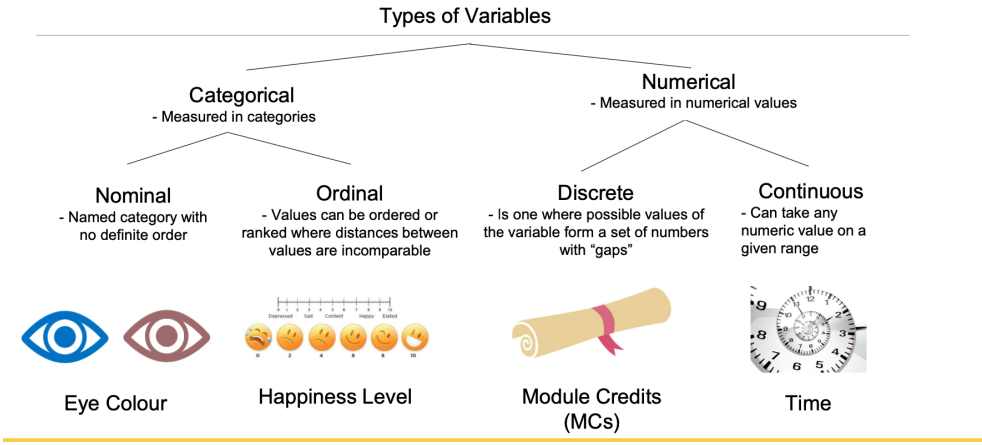
Sometimes categories come with some natural ordering and numbers are often used to represent the ordering. We call such variables **ordinal**. For example a happiness index can be rated from 0-10 in order of increasing happiness. But it is important to note that this does not make the variable numerical! (Since finding the standard deviation of happiness still makes no sense)

In other cases where there is no intrinsic ordering for the variables, we refer to these variables as **nominal**. For example if one were trying to collect basic information on a sample of birds, the eye colour (Blue / Brown) can be considered a nominal variable since there is no intrinsic ordering.

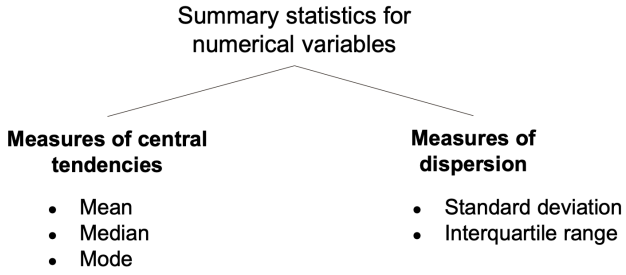
Numerical variables take numerical values for which arithmetic operations such as adding and averaging make sense. Age, measured in years for example, is a numerical variable. Mass (kg) and height (m) are also numerical variables.

Discrete numerical variable : Is one where possible values of the variable form a set of numbers with “gaps” .Example : Population count

Continuous numerical variable : Is one that can take on all possible numerical values in a given range or interval. Examples : Time, length.



▼ Summary Statistics



Mean

The **mean** of a numerical data is the "average" of all its data points. If $x_1, x_2, \ldots, x_n$ are data points for a variable x, then the mean of x, denoted by $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

Properties of Mean

1. Adding a constant value to all the data points (be it positive or negative) changes the mean by that constant value.
2. Multiplying all the values to all the data points by a constant number c will result in the mean also being multiplied by c.

Limitations of Mean

1. The mean cannot be calculated for categorical data, as the values cannot be summed.
2. As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.
3. The mean tells us nothing about the actual distribution of the data. For example, if you were told that the average income in country A is $5000 and the average income in country B is $25000, which country would you prefer to work in? B? Assume there are only 5 people in country A whose incomes are $[5000, 5000, 5000, 5000, 5000]$ and similarly in country B $[1, 1, 1, 1, 24995]$ and the last person in country B is the dictator or a drug lord. Are you sure you still want to go to country B? It is 80% likely that you will be making $1.

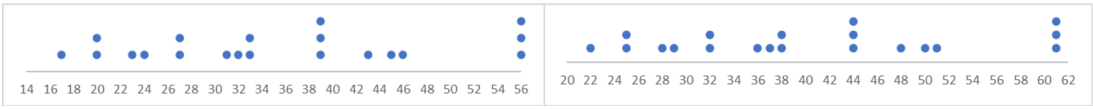
Standard Deviation

Since the mean cannot tell us anything about the "spread" of our data, we use the sample standard deviation to measure this.

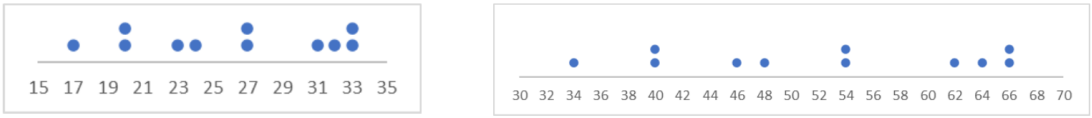
$$Sample\ Variance = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, where\ \sigma = s_x = \text{standard deviation}$$

Properties of Standard Deviation

1. It should be fairly obvious from the formula that standard deviation is always non-negative (since it is the positive square root of variance)
2. Adding a constant value c to all the data points does not affect the standard deviation. This is because shifting a set of points either left or right or up or down by a fixed amount does not affect the "spread" of the data.



3. Multiplying a constant c to all the data points causes the standard deviation to also be multiplied by a factor of $|c|$. It is necessary to take the absolute value of c since even if c is a negative constant, the standard deviation would have to be positive.



Coefficient of Variation

Given 2 data sets and their standard deviations, how would you determine which is more "spread" out?

It is not possible to just compare the standard deviations because the standard deviation can be large simply because the values themselves are large and so the square of their difference with the mean might be larger.

So, it is important to find out the "spread" relative to the mean.

This is where coefficient of variation becomes important.

If s_x is the standard deviation of x, then $coefficient\ of\ variation = \frac{s_x}{\bar{x}}$. This sort of "normalises" the standard deviation so that we can use it for comparison purposes with other data sets.

Median

Like mean, the median is important in finding out the "centre" of a distribution.

The median of a numerical variable in a data-set is the middle value of the variable after arranging the values of the data-set in ascending/descending order.

Properties of Median

1. Adding a constant value (positive or negative) to all the data points changes the median by that constant value.
2. Multiplying all the data points by a constant value c results in the median being multiplied by c.

Note that if the median of a set of 50 value is x_1 and the median of a set of another 30 values is x_2, then the median of all the 80 values is **NOT** equal to $\frac{50 \times x_1 + 30 \times x_2}{50 + 30}$ although this would be true in case of mean.

Quartiles and InterQuartile Range (IQR)

The first quartile (denoted by Q_1) is the 25^{th} percentile of the data values and the third quartile (denoted by Q_2) is the 75^{th} percentile of the data values. (25^{th} percentile simply means the median of the first half of the values when arranged in ascending order. Similarly, for 75^{th} percentile).

The practical interpretation of this is that 25% of the data points are less than or equal to Q_1 and 75% of the data points are less than or equal to Q_3.

The InterQuartile Range is the difference between Q_3 and Q_1, i.e., $IQR = Q_3 - Q_1$. It should seem obvious that 50% of the data falls within Q1 and Q3

This gives us another way to quantify the spread of the data.

Properties of IQR

1. It should be fairly obvious from the formula that IQR is always non-negative (since $Q_3 > Q_1$ under the assumption that the values are sorted in ascending order)
2. Adding a constant value c to all the data points does not affect the IQR. This is because shifting a set of points either left or right or up or down by a fixed amount does not affect the "spread" of the data.
3. Multiplying a constant c to all the data points causes the IQR to also be multiplied by a factor of $|c|$. It is necessary to take the absolute value of c since even if c is a negative constant, the IQR would have to be positive.

For this module, when we have an odd number of data points, we will not include the median in both halves if we are computing the quartiles manually.

Mean vs Median?

- Both the mean and the median can be used to describe where the "center" of a dataset is located.
- It's best to use the mean when the distribution of the data values is symmetrical and there are no clear outliers.
- It's best to use the median when the the distribution of data values is skewed or when there are clear outliers.

Mode

The value that occurs most frequently in the data set. If there are multiple such values, all of them are considered to be the mode.

Modes can be interpreted as the peak of the distribution. In terms of probability, if all the points have equal probability of being chosen, the mode has the highest probability of being chosen (simply because there are more of them)

▼ Study Designs

There are mainly 2 types of studies conducted - experimental and observational.

Experimental Studies

An experiment intentionally manipulates one variable in an attempt to cause an effect on another variable.

The primary goal of an experiment is to provide evidence for a **cause-and-effect relationship** between two variables.

There has to be a **treatment group** and a **control group** to ensure that the effect is actually **caused** by the manipulated variable and not some other confounder. **The control group provides a baseline for comparison with the treatment group**.

Random Assignment

To establish a cause-and-effect relationship, we want to make sure that the independent variable is the only factor that impacts the dependent variable. How do we account for the effects from all these other variables? Random assignment!

If the number of subjects is large, by the laws of probability, the treatment and control groups will tend to be similar in all aspects. (The law of large numbers)

The treatment and control groups can have different sizes. As long as the size of the groups are quite large, then a randomised assignment tends to produce two very similar groups.

Here, random does not mean “haphazard”. Instead, it means impartial.

Also, it is important to note the difference between random assignment and probability sampling - they are completely different!

Blinding

To prevent the participants (or researchers) from knowing which participant is in which group (treatment or control), blinding (or double blinding) is used.

Placebo: Treatment with no active ingredients, and no effect.

Placebo Effect: The response observed when subjects receive a placebo treatment, but still show some positive effects.

Blinded subjects do not know whether they are in the treatment or control group.

◦ A placebo that is very similar to the treatment can be chosen to help make the blinding effective.

◦ The subjects are blind to the treatment to prevent their own beliefs about the treatment from affecting the results.

Blinded assessors do not know whether they are assessing the treatment or control group.

An experiment is called **double-blind** if both **subjects** and **assessors** are blinded about the assignment.

Observational Studies

An observational study observes individuals and measures variables of interest. However, researchers do not attempt to directly manipulate one variable to cause an effect in another variable. We use the term exposure variable and response variable to represent the independent and dependent variables respectively.

So, **an observational study does not provide convincing evidence of a cause-and-effect relationship.**

Note: For observational studies, while there is no actual treatment being assigned to the subjects, we still use the terms “treatment” and “control” groups in the same way as though we are dealing with an experiment.

Experimental Study	Observational Study
Participants are assigned into treatment and control group by the researchers	Participants assign themselves into treatment and control groups, i.e., researchers do not control the assignment of groups
Can provide strong evidence in favor of a causal relationship between 2 variables	Can only provide evidence of association/correlation (NOT CAUSATION) between 2 variables

▼ Chapter 2

Categorical Data Analysis

Rates

Marginal rates / proportions / percentages

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

- What proportion of the total number of patients underwent Treatment Y?
- $\text{rate(Y)} = \frac{350}{1050} = \frac{1}{3} = 33\frac{1}{3}\%$
- What proportion of the total number of patients had a successful treatment?
- $\text{rate(Success)} = \frac{831}{1050} = 0.791 = 79.1\%$
- Calculations above are called marginal rates / proportions / percentages.

Conditional rates / proportions / percentages

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

- If we focus on patients who underwent Treatment X, what proportion of them had a successful treatment?
- $\text{rate(Success given X)} = \frac{542}{700} = 0.774 = 77.4\%$
- Calculation above is known as a conditional proportion / percentage.
- An even shorter way of writing this is to use a vertical bar in place of given: rate(Success | X)

Joint rates / proportions / percentages

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

- What is the proportion of patients who chose Treatment Y and had a failure?
- $\text{rate(Y and failure)} = \frac{61}{1050} = 0.0581 = 5.81\%$
- NOT a conditional rate.
- Calculation is known as a joint rate/ proportion / percentage.

Association

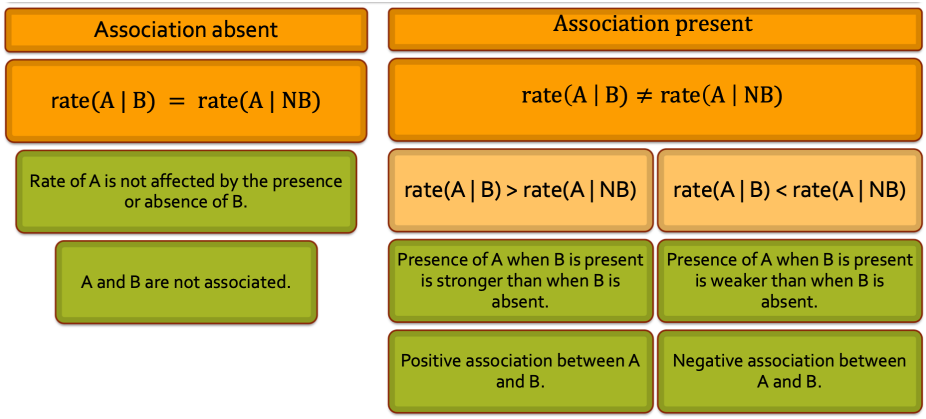
We say that X is associated with Y if knowing the state of X changes the state of Y. (It is important to remember that this does not mean that X causes Y: Corelation does not imply causation.

(Here, $\sim X$ denotes the negation of X)

Positively Associated: We say that X is positively associated to Y if $rate(Y|X) > rate(Y| \sim X)$ or $rate(X|Y) > rate(X| \sim Y)$.

Negatively Associated: We say that X is negatively associated to Y if $rate(Y|X) < rate(Y| \sim X)$ or $rate(X|Y) < rate(X| \sim Y)$

If two variables are neither positively nor negatively associated, we can say that they are not associated at all.



Symmetry rule of Rate

1. $rate(A|B) > rate(A|\sim B) \iff rate(B|A) > rate(B|\sim A)$
2. $rate(A|B) < rate(A|\sim B) \iff rate(B|A) < rate(B|\sim A)$
3. $rate(A|B) = rate(A|\sim B) \iff rate(B|A) = rate(B|\sim A)$

As a direct consequence of the symmetry rule, to check for association between two variables we only need to calculate two conditional probabilities - probability of one event but conditioned on two complementary events.

Basic rule of Rate

The overall $rate(A)$ will always lie between $rate(A|B)$ and $rate(A|\sim B)$

(this can be explained in terms of probability as follows: $P(A) = P(B)P(A|B) + P(\sim B)P(A|\sim B)$ using total probability theorem. This is simply the weighted average of $P(A|B)$ and $P(A|\sim B)$, each weighted by their respective probabilities on the condition. Obviously, the result must lie between them. Hence, the basic rule of rates is true)

1. The closer $rate(B)$ is to 100%, the closer $rate(A)$ is to $rate(A|B)$ (Use the formula above - give more weightage to the one with higher probability)
2. If $rate(B) = 50\% = rate(\sim B)$, then $rate(A) = \dfrac{rate(A|B) + rate(A|\sim B)}{2}$
3. If $rate(A|B) = rate(A|\sim B)$, then $rate(A) = rate(A|B) = rate(A|\sim B)$ (because once you factorise the $rate(A|B)$ to be common, $rate(B) + rate(\sim B) = 1$)

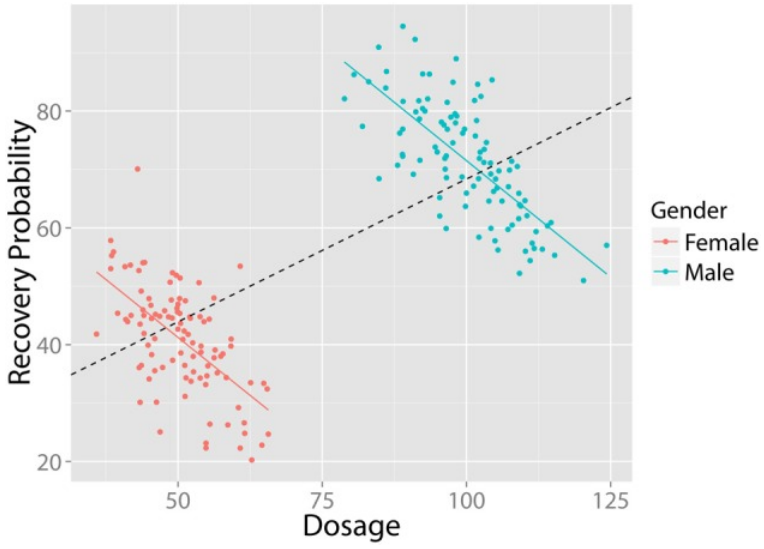
Simpson's Paradox

Simpson's paradox occurs when trends appear in different groups of data but disappear (or even reverse) when the groups are combined.

For example, consider a case where there are two possible treatments for kidney stones - X and Y.

if $rate(success|X) > rate(success|Y)$ for both large kidney stones as well as small kidney stones, but overall $rate(success|X) \leq rate(success|Y)$then this is an example of Simpson's paradox.

Another example is as follows: For both males and females, the recovery probability decreases with increase in dosage. But when both groups are considered together, the reverse trend appears.



Simpson's paradox indicates the presence of a confounder (but the converse is not necessarily true). That is,

Simpson's paradox $\implies$ **Confounder**

Confounder $\not\Rightarrow$**Simpson's Paradox**

A **confounder** is a third variable that is associated to both the variables under investigation. In the kidney stone example, the size of the stone was the confounder since it was associated with the likelihood of success and the treatment group.

In statistics, a confounder (also confounding variable, confounding factor, extraneous determinant or lurking variable) is a variable that influences both the dependent variable and independent variable, causing a spurious association. Confounding is a **causal** concept, and as such, cannot be described in terms of correlations or associations. The existence of confounders is an important quantitative explanation why coreelation does not imply causation. A confounder is the cause of the trend observed between the investigating variables

The direction of association of the confounder with the 2 variables does not matter. A confounder can have both positive associations with the variables, both negative associations or 1 positive and 1 negative.

How to solve the problem of Confounders?

There can be hundreds of possible confounders for a single experiment. Slicing is not possible for each of them and it is certainly not feasible to collect all kinds of data for privacy, ethical, time, efficiency issues. So, the best way to minimise the role of confounders is by using random assignment.

Through random assignment, we can expect that all the groups will have a similar proportion of characteristics (say, gender, age, etc.) due to pure chance.

Sometimes, randomisation is not possible due to ethical concerns (e.g. you cannot put someone under treatment X if they want to go for treatment Y instead)

▼ Chapter 3

▼ Univariate EDA

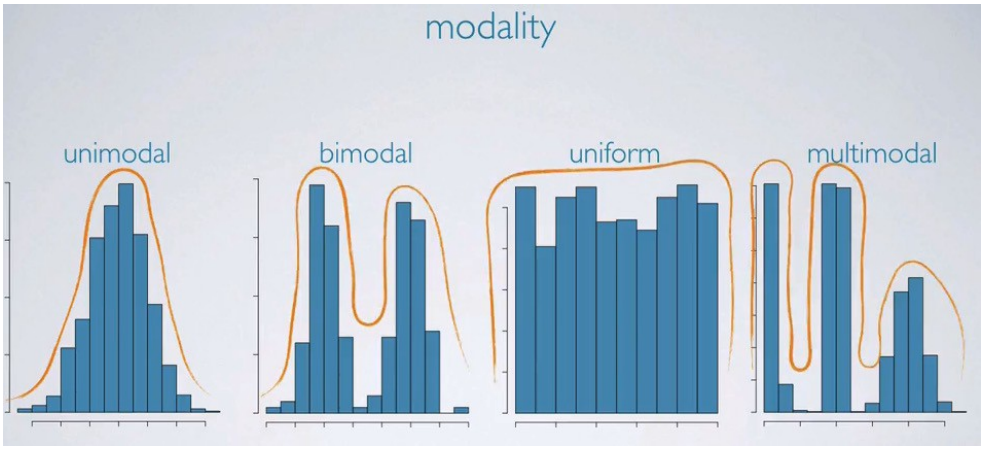
We use histogram and box-plots to gain insights about univariate data.

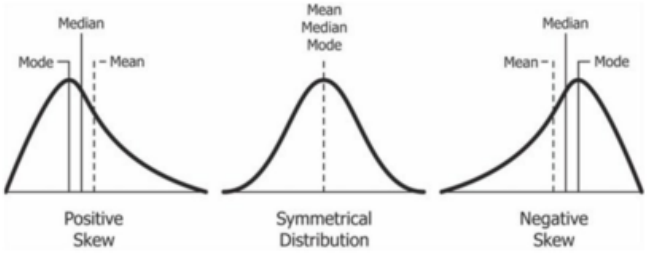
▼ Histograms

- Graphical display of a distribution
- Quick and easy to grasp
- Useful for large data sets
- Easy to obtain mode from a histogram - difficult to obtain mean and median

When descrbing distributions, we either describe the overall pattern (like shape, center and spread) or the deviations from the pattern (like outliers).

A distribution's shape is determined by 2 factors: its **peak** and its **skewness**



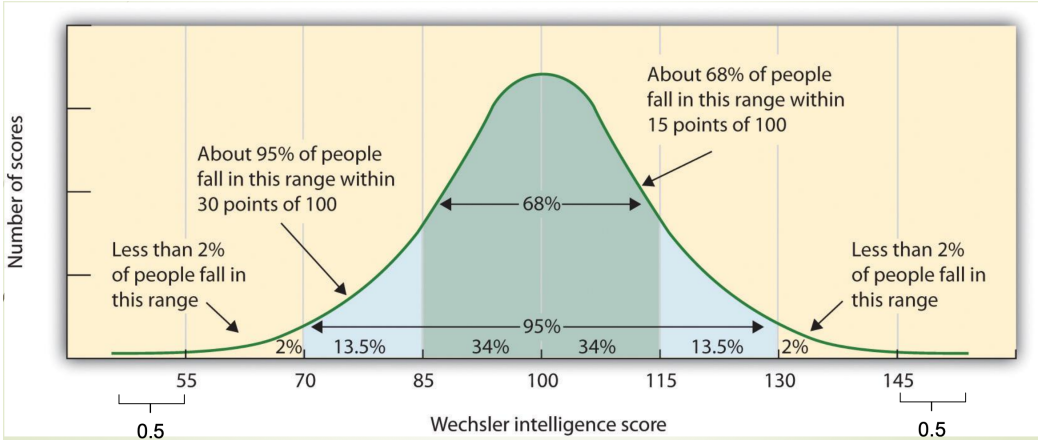


A positive skewed distribution is also called **right-skewed** (the tail of the distribution is longer on the right). Similarly a negative skewed distribution is also said to be **left-skewed** as the tail is longer on tehe left.

In case of right-skewed distribuion: Mean > Median > Mode

In case of a left-skewed distribution: Mean < Median < Mode

The folliwing is an example of a normal distribution:

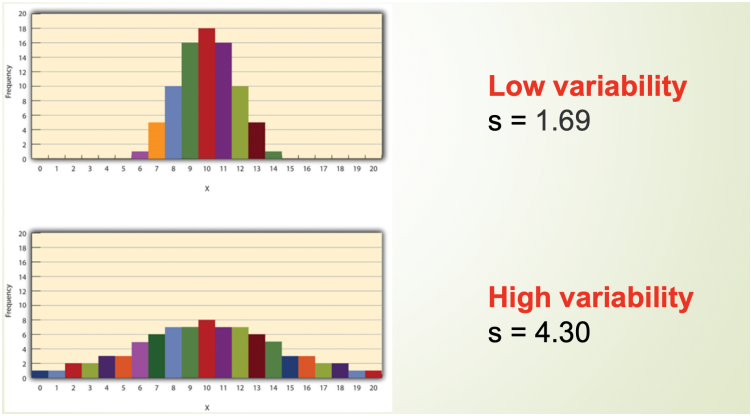


A normal distribution is always symmetric about the vertical line $x = \bar{x}$, where $\bar{x}$ denotes the mean of the distribution.

In other words, Mean = Median = Mode

Spread

The **spread** of the distribution is determined by the range and standard deviation. For example,



Outliers

Outliers are the observations that fall well above or well below the overall bulk of the data

Examining data for outliers can be useful in

- identifying strong skew in a distribution
- identifying possible data collection or data entry errors
- providing interesting insight into the data

It is often a good practice to repeat analysis with and without the outliers.

Regarding Bin Sizes of Histograms

- Avoid histograms with large bin widths that group data into only a few bins.
- Avoid histograms with very small bin widths that group data into too many bins.
- Construct histograms with different bin sizes to see which one is the most useful for our purpose.

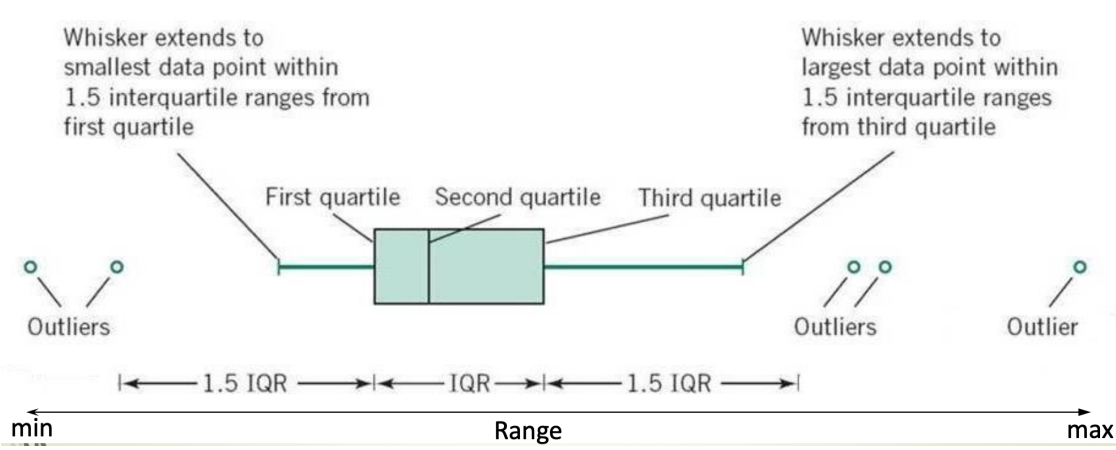
▼ Box plots

Box plots give us a lot of useful information at a quick glance. In particular, it describes the 5 number summary:

1. Minimum value
2. First quartile
3. Median (Second Quartile)
4. Third quartile
5. Maximum value

Note: A data point is considered to be an outlier if it is $> Q3 + 1.5 * IQR$ or $< Q1 - 1.5 * IQR$, where $IQR = Q3 - Q1$ is the InterQuartile Range.

Boxplots also help us identify outliers since the whiskers are exactly $1.5 * IQR$ long. That is,



▼ Bivariate EDA

Association

Natural variability exists in measurements of two variables. Average value of one variable can be described given the value of the other variable. If it is possible to predict the value of a variable based on the value of another variable, the 2 variables are said to be associated or correlated.

There are 3 main ways to check for association

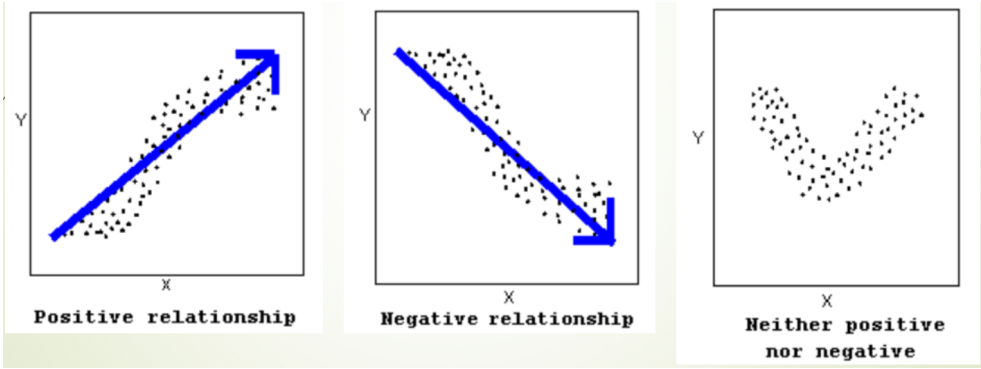
1. Scatter plots
2. Correlation coefficient - for linear relationships
3. Regression - fit a line or curve

Scatter Plots

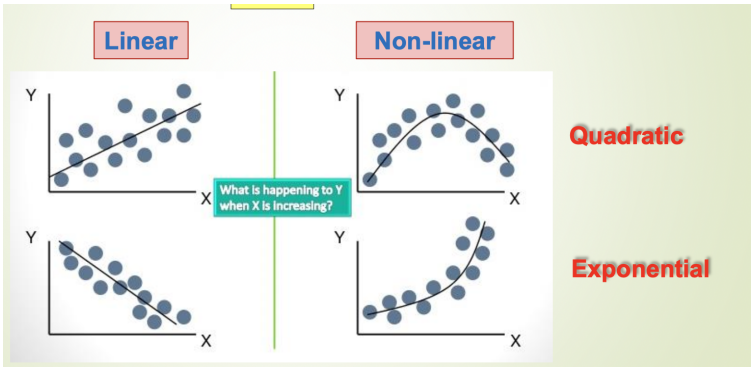
A scatter plot is plotted between the **independent (explanatory) variable** and the **dependent (response) variable.**

There are 3 main features of a scatter plot - direction, form and strength of association. Outliers can also be easily identified form the scatter plot.

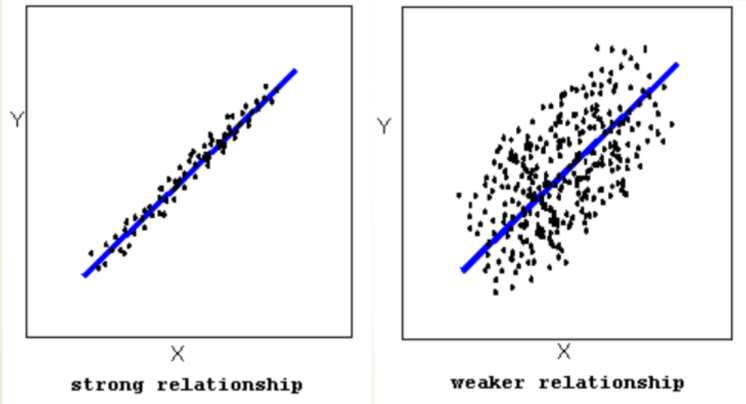
Direction refers to the nature of association - it can be either positive or negative (or neither).



The **form** describes the nature of the actual relationship (equation) between the associated variables:



Strength indicates how closely the variables are associated with each other. The stronger the association, the better you can predict the value of Y gievn the value of X.



Correlation Coefficient

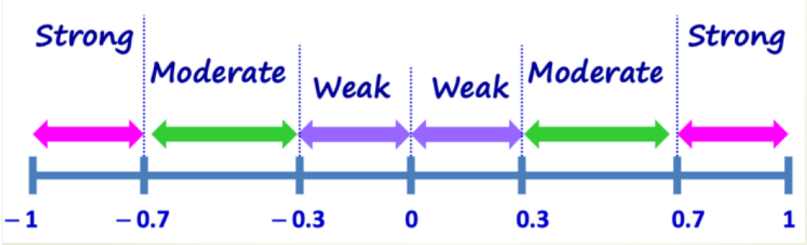
$$r = \rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_x \sigma_y}$$

- Correlation coefficient is a measure of **linear** association
- Range is between -1 and 1
- It summarizes direction and strength of linear association
  - $r > 0 \implies$ positive association
  - $r < 0 \implies$ negative association
  - $r = 0 \implies$ No **linear** association
  - $r = 1 \implies$ perfect positive association
  - $r = -1 \implies$ perfect negative association

$r = 0$ **does not imply** that there is no association between the 2 variables (or that the 2 variables are independent). It simply indicates an absence of **linear** association.

It is possible for $r = 0$ and the two variables to have a non-linear relationship (for example, quadratic)

The closer the value of $|r|$ is to 1, the stronger is the association.

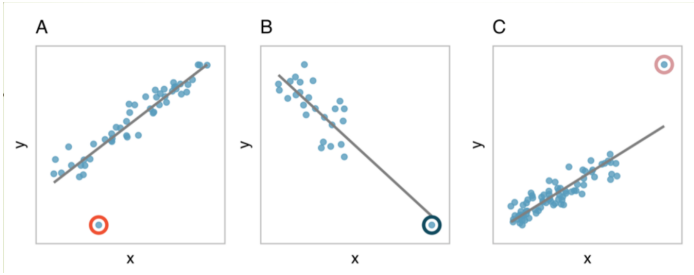


Properties about the Correlation Coefficient

- Interchanging the variables does not change the r value.
- Adding a constant to all values of a variable does not change the r $\frac{E[(X-\mu_X)(Y-\mu_Y)}{\sigma_x \sigma_y}$ constant)
- Multiplying a **positive** constant to all values of a variable does not change the r value.

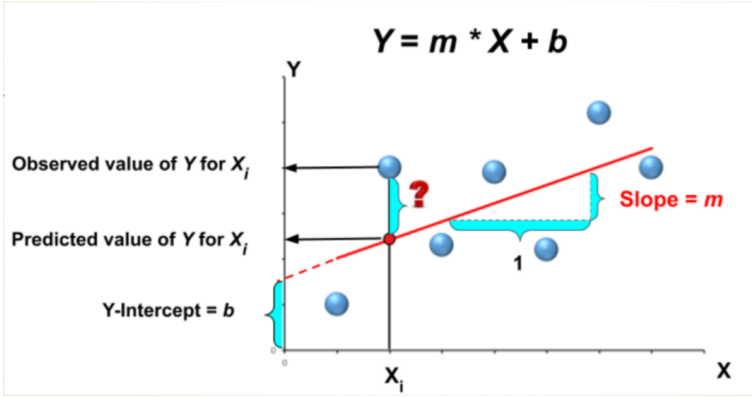
Limitations of r

- Correlation does not imply causation: even if $r = 1$, we cannot conclude that one variable **causes** the change in another variable.
- r only deals with **linear** association and not other kinds of associations. Therefore, it is necessary to first look at the scatter plot for some obvious signs of non-linear associations before calculating the correlation coefficient.
- Outliers may increase/decrease/have no effect on the correlation coefficient.



Linear Regression

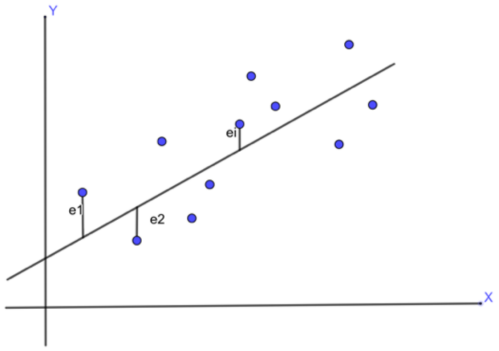
We model the relationship between 2 variables as a straight line of the form $Y = mX + b$



The regression line (like any other straight line) is completely determined by the values of m and b.

These values are calculated by minimising the residues using the least squares method. We define the i^{th} residual of the observation: e_i = difference between the observed value of Y(outcome) and the predicted value of Y(theoretical outcome).

Our aim is to minimie the sum of all the residuals: $e_1^2 + e_2^2 + \cdots + e_n^2$ where $n =$ number of data points.



It is important to remember that the regression line plotted for 2 variables depends on which is treated as the independent variable (x) and which is treated as the dependent variable (y). You cannot use the same regression line to predict the value of x, given the value of y simply because we have obtained the regression line by minimizing the deviation between theoretical and observed values for y, NOT x.

Slope of Regression Line

The slope of the regression line can be expressed in terms of the correlation coefficient as follows: $m = \frac{s_y}{s_x} r$ where s_y, s_x are the standard deviations for y and x respectively.

To get the y-intercept of the line using $\bar{x}, \bar{y}, s_x, s_y, r$ you can first find the slope m and then substitute the means as the x, y values (they should satisfy the equation).

For example, $\bar{x} = 1, \bar{y} = 2, s_x = 0.5, s_y = 2, r = 0.5$ and we are asked to find the equation of the regression line, we proceed as follows:

1. Find the slope of the regression line to be: $m = \frac{s_y}{s_x} r = 2$
2. $y - y_1 = m(x - x_1)$ be the regression line. Then, $(\bar{x}, \bar{y})$ must lie on the line. So we get, $y - 2 = 2(x - 1) \implies y = 2x$

Extrapolation

It is dangerous to extraploate the regression line beyond the observed values. The regression line can only be used accurately to predict y values when the x value is within the range of the data points collected.

Linear Regression on Non-linear Models

Say, we have a relationship between two variables that resembles an exponential curve, ie., $y = cb^t$ where y, t are the variables. (Often, it is helpful to look at the scatter plot to predict the form of the relationship)

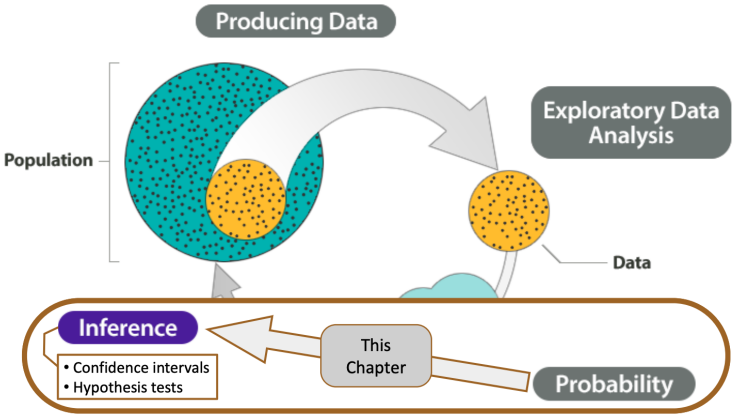
Our aim is to find b, c using the linear regression model.

We can plot $ln(y) = ln(c) + tln(b)$ as a straight line against $ln(y)$ vs. t. We can use this to find $ln(c), ln(b)$ and hence, c, b.

▼ Chapter 4

Statistical Inference

In the previous chapters, we saw ways to collect and analyze data from a sample. But, to what extent can this data be generalised to the target population? We deal with this question in this chapter



Probability

Sample Space: Collection of all outcomes of a probability experiment

Event: Subset of the sample space

All outcomes are events but not all events are outcomes!

Rules of Probability:

1. $0 \leq P(E) \leq 1$ for any event E.
2. $P(S) = 1$ where $S =$ Sample space
3. $P(E \cup F) = P(E) + P(F)$ if E and F are mutually exclusive (disjoint) events

Note: for more detailed analysis on probability and conditional probability, refer to chapter 1 of ST2334 Statistics and Probability.

Independent events are non-associated but unassociated (uncorrelated) events need not be independent

Continuous Random Variable

The probability that a continuous random variable takes on a value in the interval $[a, b]$ is equal to the area under the probability density function of the random variable from $x = a$ to $x = b$

Normal Distributions

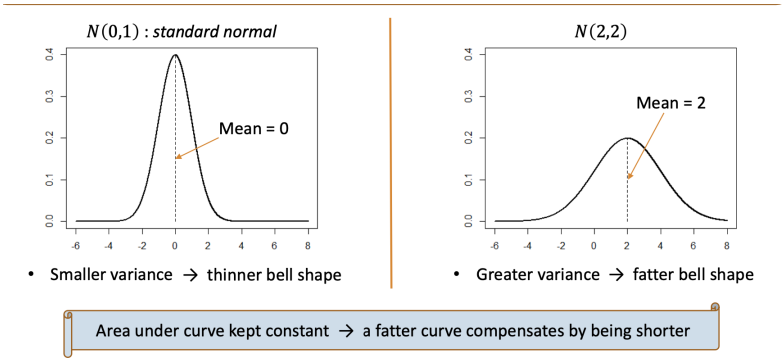
A random variable X is said to follow a normal distribution $N(\mu, \sigma^2)$ where $\mu =$ mean (expectation) of the random variable X and $\sigma =$ standard deviation.

A normal distribution is symmetric about $x = \mu$ and has a bell-curve shape.

The value of X peaks at $X = \mu$ and its value at that point is $\frac{1}{\sigma\sqrt{2\pi}}$.

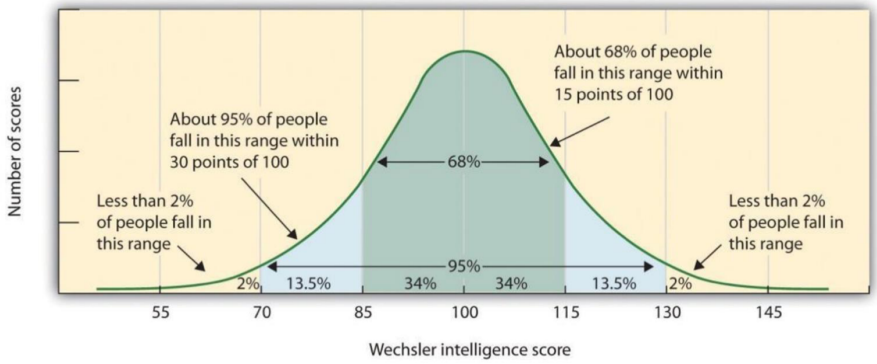
Mean = Median = Mode for any normal distribution. In other words, a normal distribution cannot be left-skewed or right-skewed.

A normal distribution is completely determined by only 2 parameters - its mean and variance. That is, 2 normal distributions can only differ in the means or variances.



Since the normal distribution is a probability density function, $\int_{-\infty}^{\infty} f(x)dx = 1$ where $f(x) = N(a,b)$ for any a, b.

A common example of a normal distribution is the distribution of IQ scores of individuals

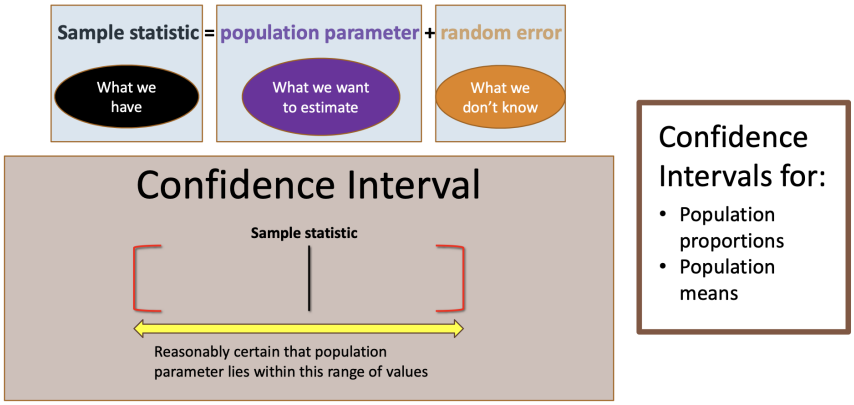


Now, we have the necessry tools to understand statistical inference.

In general, Sample statistic = population parameter + bias + random error. But we can eliminate bias by using various probability sampling techniques. So, **Sample statistic = population parameter + random error.**

There are 2 types of statistical inferences: **confidence intervals** and **hypothesis testing**

Confidence Interval



Confidence Intervals for Population Proportion

Confidence Interval for population proportion: $p^* \pm \left(z^* \times \sqrt{\frac{p^*(1-p^*)}{n}} \right)$,

where

- p^* **is the sample proportion,**
- z^* **is the value from the standard normal distribution**
- n **is the sample size.**

For example, for a 90% confidence interval, the value of z^* from the standard normal distribution is 1.645

For example, for a 95% confidence interval, the value of z^* from the standard normal distribution is 1.96

But, what do confidence intervals actually indicate?

A 95% Confidence Interval (CI) shows that we are 95% confident that the population parameter lies within this interval.

It is important to note that we use the word “confident” and not “chance”, i.e., we do not say “there is a 95% chance that the population proportion lies within this interval” because this is not true. The population parameter is a constant - it either does lie within the interval or it does not. There is no probability involved. The chance and probability is in the sampling procedure which we use to determine the parameter, not the parameter itself

So, a 95% confidence interval actually means this: If 100 samples of the same size were collected, and their respective confidence intervals calculated in the same way, then about then about 95 out of the 100 confidence intervals will contain the population parameter.

Some properties of Confidence Intervals:

- For the same sample size, to get a larger confidence in your interval, the size of your interval must increase. This is because the value of z^* increases as the level of confidence increases.
- For the same level of confidence, the interval size is smaller if your sample size is larger. (This makes intuitive sense because you have calculated the sample statistic from a larger propotion of the population and so you have a more accurate sense of the population parameter. Also, using the formula, confidence interval size is inversely related to the sample size

Clarifications regarding Confidence Intervals:

- A confidence interval of 95% does not mean that 95% of the population data lies within this interval
- A confidence interval of 95% does not mean that there is a 95% chance that the true population mean falls within the confidence interval. (The chance that the true population mean falls within the confidence interval is either 0 or 1, nothing in between)
- A confidence interval of 95% does not mean that the mean will fall within the confidence interval 95% of the time. (The mean is fixed! For one confidence interval, the mean either lies in it or not!)

Confidence Intervals for Population Means

Confidence Interval for population mean: $\bar{x} \pm \left(t^* \times \frac{s}{\sqrt{n}} \right)$,

where

- $\bar{x}$ is the sample mean
- t^* is the value from the $t-$distribution
- s is the sample standard deviation
- n is the sample size

We denote the population mean to be μ in general.

Hypothesis Testing

We are often posed with a “yes/no” question regarding data. We use hypothesis testing to answer such questions

There are a few key steps to hypothesis testing.

Step 1: Identify the question and state the null hypothesis and alternative hypothesis.

We produce 2 hypotheses namely the null hypothesis and alternative hypothesis.

The null hypothesis takes a stance of no difference or no effect. This hypothesis assumes that any differences seen are due to the variability inherent in the population and could have occurred by random chance.

Alternative hypothesis is typically what we wish to confirm and pit against the null hypothesis.

The two hypotheses must be *mutually exclusive*: they cannot both be true.

Typically, if our alternative hypothesis is that a value x is less than some value c, then our null hypothesis would be $x = c$. We don’t need to take $x \geq c$ because $x = c$ will give the maximum of the $p - values$ given by $x \geq c$. Just make sure that the null and hypothesis are mutually exclusive - try to take null hypothesis as “something equal to something” or “something is not associated with something”

Step 2: Collecting relevant data. Decide on the relevant test statistic.

The test statistic is a value computed using data which you use to determine whether to reject the null hypothesis or not.

Step 3: Determining the level of significance and computing the p-value.

The lower the significance level, the greater the evidence needs to be in order to conclude the alternative hypothesis over the null.

A commonly used level of significance is 0.05, or 5% level of significance. Other commonly used level of significance is 0.10 (10% level of significance) or 0.01 (1% level of significance)

The p-value is the probability of obtaining a test result at least as extreme as the result we observed, assuming the null hypothesis is true. The p-value can also be thought of as the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis was true.

Step 4: Making conclusion about the null hypothesis.

Now we determine whether to reject or not to reject the null hypothesis. Decide between one of the ONLY two options:

1. Reject the null hypothesis in favour of the alternative, if p-value < significance level.
2. Do not reject the null hypothesis, if p-value $\geq$ significance level. Our test result is inconclusive.

In the context of computing p-value, “at least as extreme” is interpreted as “at least as favorable to the alternative hypothesis”.

Therefore, “at least as extreme” depends on our alternative hypothesis. So, the same experiement can give us different p-values depending on our hypotheses.Thus, when making computation on p-values, it is important that we know what thenull hypothesis and alternative hypothesis is.

Common Misconceptions Clarified

- If p-value is not lower than the level of significance, we cannot reject the null hypothesis which means we don’t know if the observation is due to chance (random sampling error) or not.
- Not rejecting the null hypothesis doesn’t mean the null hypothesis is true. There does not exist a scenario where we attempt to reject alternative hypothesis. The p-value is calculated based on the null hypothesis; we can’t reject the alternative hypothesis.
- We only carry out hypothesis tests when working with sample data. When given population data, we do not conduct hypothesis tests (because we already have all the data we need to make a conclusion).
- We can make two possible types of errors - rejecting the null hypothesis when it is in fact true (called type I error - a serious kind of error) (or) not rejecting the null hypothesis even when it is false (type II error). Observe that for any decision, we can make at most one error. A rather interesting observation is that we will never know (unless otherwise proven by future tests) that our hypothesis was true or false. In particular, there are only 2 kinds of theories in the world - theories that have been proven false, and theories that are yet to be proven false. In short, even after we make a decision we dont know if the statement is true or false. We are simply making a decision based on statistical inference - NOT determining whether the null or alternative hypothesis is true.

Common Hypothesis Tests

One-Sample t-test

Chi-Squared test

It is commonly used to check whether two categorical variables, A and B are associated at the population level.

One Sample t-test	Chi-squared Test
-Mainly used when testing for significant difference between sample mean and a known/hypothesized mean	-Mainly used when testing for association between two categorical variables
-population distribution should be approximately normal if n, the sample size, is smaller than 30.	-The data given is the count for the categories of a categorical variable.
-Data used is acquired randomly.	-Data used is acquired randomly.

▼ Some Advanced Information

▼ Correlation Coefficient and Regression

To define correlation coefficient r or ρ, we first need to understand expectation, variance, covariance and standard deviation. Look at the ST2334 Notes for information on these. Once you know all that, we can define $r = \rho_{x,y} = \frac{Cov(X,Y)}{s_x s_y} = \frac{E[(X-\bar{x})(Y-\bar{y})]}{s_x s_y}$ where X and Y are random variables with mean $\bar{x}$ and $\bar{y}$, and standard deviations s_x and s_y

If $r > 0$ it means that whenever $X > \bar{x}$, it is likely for $Y > \bar{y}$. Well, how likely is it? The closer r is to 1, the greater the chance this is true. In other words, the magnitude of r tells us how closely X and Y are associated. The sign of r tells us the direction of association.

r can also be derived from the regression line between the two variables (and vice versa). If m is the slope of the regression line, then $m = \frac{s_y}{s_x} r$. In other words, the sign of m and r must be the same (since both s_y, s_x must be non-negative). Moreover, $m = 0 \ (or) \ m = \infty \iff r = 0$, i.e., the regression line is horizontal if, and only if, the correlation between X and Y is zero. To understand why this must be the case, first realize that for any constant c, $Cov(X, c) = 0$. If the regression line is perfectly horizontal, it means that for every value of x, the predicted value of y is the same. This is true only in 2 cases:

1. Y assumes only 1 value and so, $s_y = 0$. This means that Y is actually a constant, so, $Cov(X, Y) = 0$, which implies that $r = 0$.
2. For every value of x, there exist values of Y above and below the predicted value of y (that's how you draw the regression line). So, knowing the predicted value doesn't tell you whether a particular point would lie above or below the mean since there is equal chance of both occurring.

Given a regression line $Y = mX + c$, the following are always true (notice we use capital letters to indicate random variables, and small letters to indicate their realizations):

1. **$\bar{y} = m\bar{x} + c$, where $\bar{x}, \bar{y}$ denote the mean of x and y respectively. This means that the regression line must pass through $(\bar{x}, \bar{y})$, i.e., whenever X assumes its mean, Y must also assume its mean. This holds true regardless of the slope of the regression line: when X is at its mean, so is Y.**
2. **For any value of $X = x_0$, the predicted value of Y given by the regression line gives you the average value of Y when $X = x_0$, i.e., if $y = mx_0 + c$ (y is the predicted value of Y when $X = x_0$), then $E[Y|X = x_0] = y$**
3. **We said that $m = \frac{s_y}{s_x} r$. Another way to write this would be $m = \frac{s_y}{s_x} \times \frac{Cov(X, Y)}{s_y s_x} = \frac{Cov(X, Y)}{s_x^2}$.**
4. **We absolutely cannot use the same regression line to predict y-values (given x-values) and x-values (given y-values)! For each of these purposes, we need to find a new regression line, minimizing the error in prediction of that particular variable, and in general, the two equations are not the same.**
5. **We cannot extrapolate the regression line beyond the observed values of X since we cannot guarantee that X and Y follow the same kind of relationship after our last observed value. Regression lines are used to intrapolate and not extrapolate. Interpolation means to find a good estimate of Y given the realization of $X = x_0$ assuming we know the realizations of Y for values of $X < x_0$ and $X > x_0$.**

What regression really means

The linear model assumes that the relations between two variables can be summarized by a straight line.

It is customary to call the independent variable X and the dependent variable Y. The X variable is often called the predictor and Y is often called the criterion (the plural of 'criterion' is 'criteria'). It is customary to talk about the regression of Y on X, so that if we were predicting GPA from SAT we would talk about the regression of GPA on SAT.

Scores on a dependent variable can be thought of as the sum of two parts: (1) a linear function of an independent variable, and (2) random error. In symbols, we have: $Y_i = mX_i + c + e_i$ where Y_i is a score on the dependent variable for the ith person, $mX_i + c$ describes a linear function relating Y_i and X_i and e_i is an error. Note that there is a separate score for each X, Y, e (these are variables) but there is only one value of m and c.

Here, c denotes the x-intercept, m denotes the slope of the regression line.

The problem is that we don't know the value of each e_i. So, if we take out the error term in the equation, we have a straight line that can be used to predict the values of Y from the values of X, which is the main purpose of regression. It will look something like: $Y' = mX + c$. Notice that we use Y' here since it is a predicted value, NOT the actual value Y.

We call the difference between the actual value of Y and the value of Y' obtained using the regression line, the residual. So, when we construct a regression line, we should minimise the residuals, which can also be viewed as the error.

Finding the correct values of m and c to minimise the residuals is not easy in general (uses a lot of machine learning and algorithms such as stochastic gradient descent).

Two important things to note about Y and Y' once you find the right regression line:

1. $E[Y'] = E[Y]$, i.e., the mean of the predicted values of Y is equal to the mean of the actual values of Y. It necessarily follows then that the $E[Y - Y'] = 0$ which means that the mean of the residual values e is equal to 0.
2. The variance of Y is equal to the variance of the predicted values of Y plus the variance of the residuals.

An interesting feature is that when X and Y are expressed as their $z - scores$, the slope of the regression line is equal to the correlation coefficient. Recall that r is the average of the cross products. So, $r = \frac{\sum z_X z_Y}{N}$. A way to view this is that if $r = 1$, when $X = x_0$ is increased to $X = x_0 + s_x$, then the predicted value of Y also increases by one standard deviation, i.e., s_y. So, we can write, in general, $m = r\frac{s_y}{s_x}$ (rise over run - slope is basically the rate of change in Y per unit change in X.)

To find the intercept, $c = \bar{y} - m\bar{x}$ since $(\bar{x}, \bar{y})$ must lie on the regression line. Think about it this way, if there is no relationship between X and Y, then the best guess for all values of X is simply the mean of Y. If there is a relationship ($m \neq 0$), then still the

best guess for Y when $X = \bar{x}$ should be $\bar{y}$. As X departs from its mean, we expect Y to also depart from its mean. This is why $(\bar{x}, \bar{y})$ must lie on the line.

To summarise neatly, if y_i is the predicted value of Y when $X = x_i$, we are not concluding that $Y = y_i$. Rather, we are claiming that the average value of Y when $X = x_i$ is y_i

Least Squares

It is not enough to make each residuals as small as possible (because there is a tradeoff involved: when you fit the line such that e_i is small, you might end up making e_j bigger) - that is not a good solution. A better regression line would minimise the sum of the squares of the residuals - called as mean square errors (MSEs). This is where machine learning comes in and helps us to find the best regression line by changing the values of m and c so as to best "fit" the data. This has been proven to be the best regression line.

Least squares is called a loss function (for badness of fit or errors). It is not the only loss function in use. The loss function most often used by statisticians other than least squares is called maximum likelihood. Least squares is a good choice for regression lines because is has been proved that least squares provides estimates that are BLUE, that is, Best (minimum variance) Linear Unbiased Estimates of the regression line. Maximum likelihood estimates are consistent; they become less and less unbiased as the sample size increases. You will see maximum likelihood (rather than least squares) used in many multivariate applications.

▼ Odds Ratio

Odds is another way of describing the rate of an event. It gives the ratio of rate of occurrence of the event to the rate of non-occurrence of the event. Let $odds(A|B)$ be the odds of A among people with characterstic B. Then,

$$odds(A|B) = \frac{rate(A|B)}{rate(notA|B)} = \frac{rate(A|B)}{1 - rate(A|B)}$$

Observe that in case of odds, the condition on the two populations must be the same.

The odds ratio of A between people with B and people with "not B" (comparing odds across two different groups of people)is:

$$\frac{odds(A|B)}{odds(A|notB)} = \frac{rate(A)/rate(notA|B)}{rate(A)/rate(notA|notB)}$$

A more intuitive explanation about odds can be descibed as such: If the probability of raining on a particular day is $2/3$, we deduce that if we simulate this many times then on average, out of 3 times, it will rain 2 times. Another way to think about it is that it is twice as likely to rain than it is to not rain: here rather than comparing with the total number of times the experiment is repeated, we are comparing with the number of times the unfavourable outcome takes place. So, we can say that the odds of rain are $2:1$ or simply $2/1 = 2$.

It should be clear that if the odds are equal to 1, the probability of the favourable outcome and unfavourable outcome are both 0.5. Further, when the odds are less than 1, the probability of the favourable outcome is less than 0.5, and when odds are greater than 1, probability of favourable outcome is greater than 0.5

It is easy to convert between probability and odds once you understand the intuition behind it.