GEA1000 – Consolidated Notes

Quantitative reasoning with data (National University of Singapore)

The population is the entire group (of individuals or objects) that we are wanting to investigate or know something about. A population parameter is a ***numerical factor and a constant.***

As we rarely have access to the entire population of users, we instead, rely on a subset of the population to use as a proxy for the population. This is known as a sample (or, a proportion of the population). This method is more efficient and cost-effective than a census which is an ***attempt*** to reach out to the entire population.

Sample statistics estimate unknown population parameters. To do so, one must identify a sufficient sampling frame from an analysis of the research question.

A research question is one that seeks to investigate a particular characteristic of a population. They can:

1. *Make an estimate about the population*
   Eg. What is the average number of hours students study each week?
2. *Test a claim about the population*
   Eg., Does the majority of students qualify for student loans? And;
3. *Compare two subpopulations and investigate a relationship between two variables*
   Eg., Are student-athletes more likely than non-athletes to do final year projects?

The Exploratory Data Analysis (EDA) model relies on data selected from a portion of the population (sampling) in order to provide a reasonable estimate. ***This estimate is an inference about the population parameter based on the information gathered from a sample.***

More explicitly, EDA is a ***systematic process where we explore data and summarise it using graphs and different numerical analysis methods*** like percentages or averages to evaluate the usefulness of existing data and to refine the analysis process.

A sampling frame may not cover the population of interest or may contain units that are not in the population of interest. ***A good sampling frame should be equal to or greater than the population of interest.***

Ideally, you should select your sample randomly from the parent population but this may be difficult to do due to:

1. *issues establishing a truly random selection scheme*;
2. *problems getting selected units to participate*.

Representativeness is more important than randomness.

Bias is a source of systematic error and enters into studies in two primary ways:

1. *During the selection and retention of the subjects of the study;*

Selection Bias: Associated with the researcher's biased selection of units. The sample is selected in a way that systematically excludes part of the population; this means that the results of the study will be biased towards those who are in the sampling frame. ***Selection bias is the result of an imperfect sampling scheme or a non-probability sampling scheme.***

Eg. Volunteer Bias: the fact that people who volunteer are usually not representative of the population as a whole.

Non-Response Bias: Associated with the participants' non-disclosure of information related to the study. Just as people who volunteer to take part, those who decline also differ systematically from the rest of the population. ***Non-response bias is the result of disinterest, inconvenience or an unwillingness to disclose or discuss sensitive information.***

Eg. Informative Censoring: particularly important in longitudinal studies (a study where individuals are followed for a period of time). Although losing subjects is common, the issue lies in when subjects do not drop out at random but rather due to reasons related to the study's purpose.

This may occur regardless of probability or non-probability sampling schemes

2. *In the method in which information is collected about the subjects. This leads to* information bias.

Interview bias: When a bias is introduced into the data, from the outset, due to the attitudes or behaviour of the interviewer.

Recall bias: Especially relevant in studies investigating illnesses, people with a lifelong condition such as suffering from a serious illness are more likely to remember events that they relate to those experiences.

Detection bias: In keeping, certain characteristics are more likely to be detected in some individuals than others (pain thresholds, symptoms of depression etc.)

Social desirability bias: Caused by people's desire to present themselves in a favourable light (especially relevant to age-related studies, involvement of culture, sensitivity and reception to mental health etc.)

If the sample is biased, it is not representative of the population and the conclusion cannot be applied to the entire population. This is known as poor generalisability.

Probability Sampling employs a known randomised mechanism with the purpose of preserving the element of chance in the selection process to eliminate biases associated with selection.

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

The random selection words on the basis of selection with replacement.

- This means that there is a chance that the same unit is selected more than once. To circumvent this limitation, the researcher might choose to increase the population size or use a smaller sample in proportion to the population size.

Simple Random Sampling: Units are selected randomly from a sampling frame using a random number generator (RNG), randomly picking individuals associated with the number obtained through RNG. ***There must be a non-zero chance for each unit.***

- Advantage: Results do not change haphazardly from sample to sample. Variability is entirely due to chance. Therefore, the sample tends to be a good representation of the population as diversity is preserved.
- Disadvantage: Subject to non-response bias and accessibility of information.

Systematic Sampling: A method of selecting units from a list by applying a selection interval 'k' to a ***randomly designated starting point*** (Arithmetic Sequence).

- Advantage: Simpler selection process than simple random sampling.
- Disadvantage: It may not be representative of the population if the list is non-random.

Stratified Sampling: Breaks apart each population into different strata which may vary in size but are relatively similar in nature before randomly selecting via simple random sampling. ***The same 'x' number of units are selected per stratum.***

- Advantage: Able to get a representation sample from each stratum.
- Disadvantage: Requires information about the sampling frame and strata beforehand for classification into stratum may be performed.

Cluster Sampling: Breaks apart the population into clusters and randomly selects a fixed number of clusters, ***including all observations and units from selected clusters.***

- Advantage: Less tedious, less time consuming and less costly.
- Disadvantage: High variability due to dissimilarities in clusters or small numbers of clusters. Therefore it might not be representative of the population.

Non-Probability Sampling is the selection of individuals/units done by human discretion. ***These sampling schemes are not influenced by chance (heavily problematic).***

Convenience Sampling: A non-probability sampling scheme in which the researchers use the subjects that are most readily and easily available and willing to participate in the research study.

- Disadvantage: Highly problematic and subject to a litany of biases discussed prior, including both selection and non-response biases (heavily influenced by demographic).

Volunteer Sampling: A non-probability sampling scheme in which researchers' actively seek individuals to participate in the study.

- Disadvantage: Volunteer sampling is highly susceptible to bias because researchers make little effort to control sample composition. The people who volunteer may believe very differently from those who do not. ***This leads to systematic exclusion and a bias in the results towards those who did participate.***

	General Approach to Sampling	Generalisability Criteria
1	Choose Sampling Frame	Good Sampling Frame (Larger or Equal to Population)
2	Sample from Sampling Frame	Probability-Based Sampling Scheme
3	Remove unwanted units	Large Sample Size (Reduce Error)
4	[redacted]	Minimal Non-Response

A Data Set consists of individuals and variables pertaining to the individual. Variables are an attribute that can be measured or labelled.

Independent variables are variables that can be manipulated either deliberately or spontaneously in a study. ***"Deliberately" refers to manipulated changes made by the researchers while "spontaneously" refers to manipulations to the variable outside of the researchers' control.*** They may sometimes be known as explanatory variables.

Dependent Variables are variables that are hypothesised to change depending on how the independent variable is manipulated. They may sometimes be known as response variables.

Variables mainly take 2 forms:

1. Categorical Variables: take categories or label values that ***must be mutually exclusive.***

    a. Categories with some natural ordering, like a happiness index or Likert Scale, are known as Ordinal variables. However, in the example of the happiness index, we cannot assume that the difference between 6 & 7 is the same as 7 & 8. ***Therefore, performing arithmetic analysis, like averaging, is unadvisable.***

    b. In other cases where there is no intrinsic ordering, like eye colour, these categorical variables are known as nominal.

2. Numerical Variables: takes numerical values for which arithmetic operations (BODMAS) can be performed soundly.

    a. A discrete variable is one where possible values for the variable fall into integers and form a set of numbers with "gaps" between them (ie., population count; where numbers with decimal places don't make sense.) ***Any numerical value that can only take a finite number of possible values is automatically a discrete variable.***

    b. A continuous variable is one that can take all possible numerical values in a given range or interval (time, height, mass; where decimal outputs make sense.)

Measures of Central Tendencies.

1. Mean (Average)
   a. Can be calculated for any set of numerical data.
   b. A set of numerical data has only one mean value.
   c. The mean is affected by unusually large or small data values.
   d. Adding a constant value to all the data points (be it positive or negative) changes the mean by that constant value.
   e. Multiplying all the values to all the data points by a constant number c will result in the mean also being multiplied by c

The ***mean is used when the data is scaled*** (data with equal intervals like speed, weight, height, temperature; numerical and arithmetically) ***and the distribution is normal*** (the mean is sensitive to outliers that are found in skewed distributions). The mean cannot tell us much about the ***nature of the distribution of the data.***

2. Median (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.)
   a. There is a unique median for each data set.
   b. It is not affected by extremely large or small values; it is useful when there are extreme outliers.
   c. It is not applicable to qualitative data.
   d. Is used when the data is ordinal.
   e. Adding a constant value (positive or negative) to all the data points changes the median by that constant value.
   f. Multiplying all the data points by a constant value c results in the median being multiplied by c.

The median is ***used when the data is ordinal and the distribution is skewed or non-normal.*** The median of a numerical variable ***does not tell us the total value, frequency of occurrence or the distribution of data points of the numerical variable.***

3. Mode (a variable is the value of the variable that appears the most frequently.)
   a. The mode is not always unique; a data set can have no or more than one mode.
   b. Can be used for qualitative and quantitative data.
   c. Not affected by extreme values.
   d. In the context of probability, a peak of the distribution refers to the value that has the highest probability of occurring.

The ***mode is used when you want to know the most frequent response, number or observation in a distribution.*** The mode ***can also be used when the data is nominal or categorical*** such as religious preference, gender, or political affiliation.

Measures of Dispersion

1. Standard Deviation (a measure of spread around the mean)
    a. The standard deviation is always non-negative.
    b. Adding a constant value, c (positive or negative) to all the data points does not change the standard deviation.
    c. Multiplying all the data points by a constant value c results in the standard deviation being multiplied by c where c is the absolute value of c.

2. Sample Variance
    a. If a constant c is added to each value of a population function, then the new variance is the same as that of the old variance.
    b. If each data item of a population function is multiplied by a constant c, the new variance is c^2 times the old variance.

3. Interquartile Range
    a. Q3 - Q1 = IQR
    b. A small IQR value means that the middle 50% of data values have a narrow spread whilst a large IQR value indicates a large spread for the middle 50% of the values.
    c. The IQR is always non-negative and this follows from the fact that $Q3$ is at least as large as $Q1$.
    d. Adding a constant value, c (positive or negative) to all the data points does not change the IQR.
    e. Multiplying all the data points by a constant value c results in the IQR being multiplied by $|c|$.

4. Coefficient of Variation (way of quantifying the degree of spread relative to the mean; also known as the relative standard deviation).
    a. Calculated by standard deviation divided by the mean.

Association Absent

$Rate\ (A\,|\,B)\ =\ Rate\ (A\,|\,NB)$

- The rate of A is not affected by the presence or absence of B.
- A and B are not associated.

Association Present

$Rate\ (A\,|\,B)\ =/=\ Rate\ (A\,|\,NB)$

1. When,
   $Rate\ (A\,|\,B)\ >\ Rate\ (A\,|\,NB)$
   - The presence of A when B is present is stronger than when B is absent.
   - There is a positive association between A and B.

2. When,
   $Rate\ (A\,|\,B)\ <\ Rate\ (A\,|\,NB)$
   - The presence of A when B is present is weaker than when B is absent.
   - There is a negative association between A and B.

Basic Rule on Rates

The overall Rate (A) will always lie between Rate (A | B) and Rate (A | NB).

Consequence of the Basic Rule on Rates

1. The closer Rate (B) is to 100%, the closer Rate (A) is to Rate (A | B)

2. If Rate (B) = 50%, then $[Rate\ (A\,|\,B)\ +\ Rate\ (A\,|\,NB)]\ /\ 2$

3. If Rate (A | B) = Rate (A | NB), then Rate (A) = Rate (A | B) = Rate (A | NB)

Simpson's Paradox

It is a phenomenon in which a trend appears in several groups of data but disappears or reverses when the groups are combined. ***In order words, it is the phenomenon whereby the direction of association gets reversed when the groups are combined.*** The only (sure-fire) way to "solve" a Simpson's Paradox is through a subgroup analysis called "Slicing".

$Simpson's\ Paradox\ \rightarrow\ Confounder$ but $Counfounder\ =/=\ Simpson's\ Paradox$

When a Simpson's Paradox occurs, it implies that there is definitely a confounding variable present. However, this does not mean that a confounder necessarily leads to a Simpson's Paradox.

Confounders

<u>A confounder is a third variable that is associated with both the independent and dependent variables whose relationship we are investigating.</u>

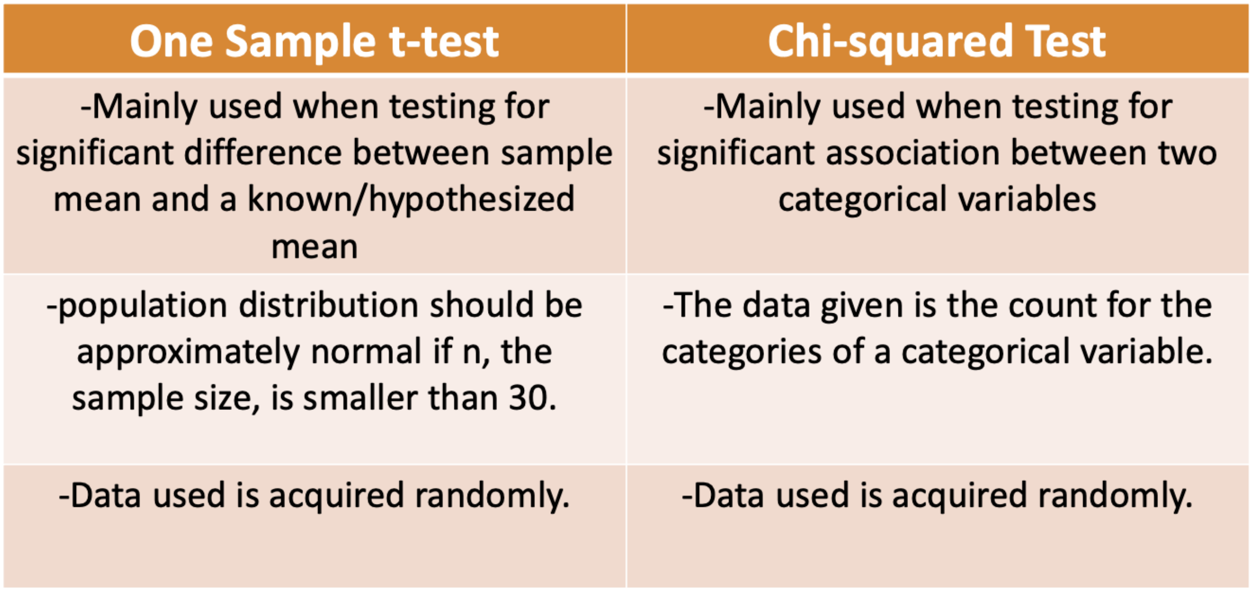
We do not specify the direction of association. As long as the variable is associated in some way with the main variables, we will call it a confounder.

Instead of trying to check for confounders one by one, randomised assignment attempts to work as a general solution across all confounders. Random assignment tends to give us equal proportions across the two groups, thus making it a good solution, especially for confounders that we did not measure. However, there are still times where we may be unlucky and end up with a disproportionate allocation even after random assignment.

Randomisation is not always possible in studies. In the kidney stone example, along with other medical trials, it is difficult to force patients to undergo a treatment that they are unwilling to do and this makes the assignment process non-random, or in this case an observational study.

Probability Experiments must be ***<u>repeatable</u>*** and must ***<u>give rise to a precise set of outcomes.</u>*** They form the bedrock of probability. Conventionally, probabilities are numerical values between 0 and 1 that are assigned to events.

A Sample Space is the collection of all outcomes of a probability experiment while an Event is a subcollection of the sample space.



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 significant 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.

Representing Data

A distribution is an orientation of data points, broken down by their observed frequency of occurrence.

Histograms are a graphical display of a distribution in a quick and easy way to grasp that can be useful for large data sets. ***To create a histogram, divide the variable values into equal-sized intervals called bins.***

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.

A unimodal distribution has 1 distinct peak. Whereas a multimodal distribution has more than one distinct peak. In a symmetrical distribution, its left and right halves are mirror images of each other and the peak is in the middle.

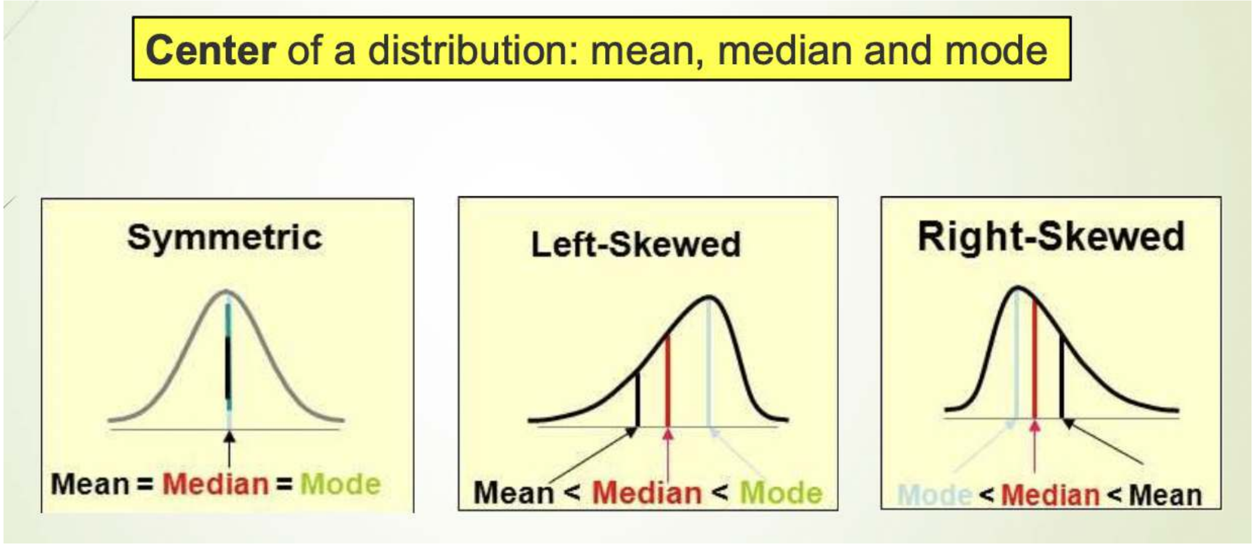
If the peak is towards the right, it is left-skewed.

If the peak is towards the left, it is right-skewed.

The Normal Distribution is a symmetrical distribution. It is unimodal where mean, median and mode and equal.

68% of the population fall between 1 standard deviation

95% fall between 2 standard deviations from the mean (distinct peak).



After graphically representing the distribution of a quantitative variable, we can describe the pattern in terms of shape (peaks and skewness), centre (mean, median mode), spread (range and standard deviation) and deviations (outliers; observations that fall well above or below the overall bulk of data) from the pattern.

Spread refers to the degree of variability.

Low variability indicates that the data is relatively close to the centre whereas the high variability indicates that the data is spread across a much wider range.

Differences between Histograms and Bar Graphs:

1. A histogram shows the distribution of a numerical variable across a number line but a bar graph makes comparisons across categories of a variable.

2. The ordering of bars in histograms cannot be changed unlike those of a bar graph.

3. Usually, there are no gaps between bars in a histogram.

Boxplots are constructed through the "Five-Number Summary": Minimum, Quartile (Q1), Median (Q2), Quartile (Q3), and maximum. (IQR = Q3 - Q1, quantifies the spread of the data around the median)

A Data Point is considered an outlier if it satisfies either one of the following conditions:

- Its value is greater than Q3 + 1.5 x IQR,

- Its value is less than Q1 - 1.5 x IQR

Boxplots vs Histograms:

1. Histograms can provide a better sense of the shape of the distribution of a variable, especially when there are great differences among the frequencies of the data points.

2. Boxplots are more useful when comparing distributions of different data sets.

3. Boxplots can identify and exhibit outliers clearly.

4. Boxplots do not give any information about how many data points we are working with. In order words, two boxplots can look the same but correspond to data sets with very different numbers of data points.

Statistical Relationship (non-deterministic relationship) between two variables where variability exists in their measurements. Unlike the case of deterministic relationships, it is not possible to find a unique value of one variable corresponding to each value of the other variable. Instead, ***we describe statistical relationships in terms of the average value of one variable in terms of the other variable.***

Bivariate Data Analysis can be performed with:

1. Scatter Plots (get an idea of the pattern)
    a. If it is not clear which variable may affect the other, we can arbitrarily choose one of them as the x variable and the other as the y variable.
        i. Direction:
            1. A positive (or increasing) relationship means that an increase in one of the variables is associated with an increase in the other.
            2. A negative (or decreasing) relationship means that an increase in one of the variables is associated with a decrease in the other.
        ii. Shape
            1. Linear: The data points appear scattered about a line.
            2. Non-Linear: The data points appear scattered about a smooth curve (quadratic/exponential/polynomials).
        iii. Strength
            1. The strength of the relationship is a description of how closely the data follows the form of the relationship.
            2. The closer the dots (data points) are to the ***line of best fit***, the stronger the ***linear relationship.***

2. Regression Analysis (fit a line or curve to the data).
    a. $y = mx + c$
        i. The regression line obtained can only be used to predict values in the domain (x-values) it covers. Beyond this domain, any prediction is dangerous as the best-fit regression line may change.
        ii. You cannot inverse the graph to predict the true is same ie., HDB price vs Age, cannot be reversed to Age vs HDB price if the original regression line was plot for HDB price vs Age.
        iii. The slope of the regression line and the correlation coefficient is related by the formula $m = \frac{Sy}{Sx}r$ where S_y is the standard deviation for y and S_x is the standard deviation for x.
            1. If the correlation coefficient is positive, then the gradient is also positive.
            2. If the correlation is negative, then the gradient is also negative.
            3. The slope of the regression line need not be equal to the correlation coefficient.
        iv. To model exponential equations, simply plot for log y against x and then solve for y from the linear equation derived.

3. Correlation Coefficients (the measure of linear association, range is between -1 and 1 and it summarises the direction and strength of linear association)
    a. r > 0, positive association
    b. r < 0, negative association
    c. ***r = 0, no linear association***
    d. r = 1, perfect positive association
    e. r = -1, perfect negative association

$$-1 < strong < -0.7 < moderate < -0.3 < weak < 0 < weak < 0.3 < moderate < 0.7 <$$

If the data points are connected by a vertical line or a horizontal line, r = 0 since the reason they are connected by a vertical or a horizontal line means that a change of value in one of the variables does not cause a change in the other.

r is not affected by: an interchange of two variables, an addition of a constant 'c' to all values of a variable, a multiplication of a positive number to all values of a variable.

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} \;,\quad where:$$

$$SS_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$SS_y = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

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

Remember: Correlation does not imply causation as there might be a third variable correlated to the two variables. ***Even if, there is a strong linear association observed (a value close to 1 or - 1) we can only conclude that there is a statistical relationship between the variables, but not a casual relationship.*** Furthermore, r only measures linear association between two variables and, in small data sets, extreme outliers in the data set will affect the r-value greatly.

Rates = Probability so they share the same formulaic relationships:

Rules of Symmetry

Positive Association

$$Rate\ (A\,|\,B)\ >\ Rate\ (A\,|\,NB)\ \Leftrightarrow Rate\ (B\,|\,A)\ >\ Rate\ (B\,|\,NA)$$

- $Rate\ (A\,|\,B)\ >\ Rate\ (A\,|\,NB)\ \rightarrow Rate\ (B\,|\,A)\ >\ Rate\ (B\,|\,NA)$
  - If the rate of A given B is more than the rate of A given NB, then it implies that the rate of B given A is more than the rate of B given NA.
- $Rate\ (B\,|\,A)\ >\ Rate\ (B\,|\,NA)\ \rightarrow Rate\ (A\,|\,B)\ >\ Rate\ (A\,|\,NB)$
  - If the rate of B given A is more than the rate of B given NA, then it implies that the rate of A given B is more than the rate of A given NB.

Negative Association

$$Rate\ (A\,|\,B)\ <\ Rate\ (A\,|\,NB)\ \Leftrightarrow Rate\ (B\,|\,A)\ <\ Rate\ (B\,|\,NA)$$

- $Rate\ (A\,|\,B)\ <\ Rate\ (A\,|\,NB)\ \rightarrow Rate\ (B\,|\,A)\ <\ Rate\ (B\,|\,NA)$
  - If the rate of A given B is less than the rate of A given NB, then it implies that the rate of B given A is less than the rate of B given NA.
- $Rate\ (B\,|\,A)\ <\ Rate\ (B\,|\,NA)\ \rightarrow Rate\ (A\,|\,B)\ <\ Rate\ (A\,|\,NB)$
  - If the rate of B given A is less than the rate of B given NA, then it implies that the rate of A given B is less than the rate of A given NB.

Equal/No Association

$$Rate\ (A\,|\,B)\ =\ Rate\ (A\,|\,NB)\ \Leftrightarrow Rate\ (B\,|\,A)\ =\ Rate\ (B\,|\,NA)$$

- $Rate\ (A\,|\,B)\ =\ Rate\ (A\,|\,NB)\ \rightarrow Rate\ (B\,|\,A)\ =\ Rate\ (B\,|\,NA)$
  - If the rate of A given B is equal to the rate of A given NB, then it implies that the rate of B given A is equal to the rate of B given NA.
- $Rate\ (B\,|\,A)\ =\ Rate\ (B\,|\,NA)\ \rightarrow Rate\ (A\,|\,B)\ =\ Rate\ (A\,|\,NB)$
  - If the rate of B given A is equal to the rate of B given NA, then it implies that the rate of A given B is equal to the rate of A given NB.

Bayes Theorem $P(B\,|\,A)\ =\ \frac{P(B)P(A\,|B)}{P(B)P(A\,|B)\,+\,P(B')P(A\,|\,B')}$

Complementary Events $P(A)\ +\ P(A')\ =1$

Combined Events $P(A\ \cup\ B)\ =\ P(A)\ +\ P(B)\ -\ P(A\ \cap\ B)$

Mutually Exclusive Events $P(A\ \cup\ B)\ =\ P(A)\ +\ P(B)$

Conditional Probability $P(A\,|\,B)\ =\frac{P(A\cap B)}{P(B)}$

Independent Events $P(A\ \cap\ B)\ =\ P(A)P(B)$