

**GEA1000 notes by @yesclaws :D**

**Guys i got an A from this so just read and use, DONT bother with their long ass long winded notes**

**Also do check out my mod review blog!**

<https://yesclawsnusmodulereview.wordpress.com/>

## **Chapter 1**

### Definitions

Population	The entire group (of individuals or objects) that we want to know something about
The population of interest	A group in which the researcher has an interest in drawing conclusions of the study <ul style="list-style-type: none"><li>- Such as 'the population of Asia', 'the population of Singapore' etc.</li></ul>
Population parameter	A numerical fact about a population <ul style="list-style-type: none"><li>- These are constants</li></ul>
Sample	A proportion of the population selected in the study <ul style="list-style-type: none"><li>- Done when census data is not readily available</li><li>- Preferred over the census as a sample is less costly administratively and collection+processing of data is faster for a sample</li></ul>
Census	An attempt to reach out to the entire population of interest <ul style="list-style-type: none"><li>- Note that one might not achieve a 100% response rate</li></ul>
Estimate	Inference about the population's parameter, based on information obtained from a sample
Sampling frame	<p>'Source material' from which sample is drawn</p> <ul style="list-style-type: none"><li>- May not cover the population of interest or may contain units that are not in the population of interest</li></ul> <p>Note: In order to fulfill generalisability criteria, the sampling frame should be equals to or larger than the target population (if the sampling frame fails to cover any member of the target population, it cannot be used to generalize fully to the target population as there exist members that have been left out)</p>

### Types of research questions:

1. Making 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?
2. Test a claim about the population
  - Does the majority of students qualify for student loans?
  - Is the average course load for a university student greater than 20 units?
3. Compare 2 sub-populations/Investigate a relationship between 2 variables in the population
  - In university X, do female students have a higher GPA score than male students (1)
  - Does drinking coffee help students pass the math exam (2)

### Sampling methods - Probability sampling

- The selection process is via a known/randomized mechanism in which every unit in the population has a non-zero and known probability of being selected
- Eliminate biases associated with human selection

#### 1. Simple random sampling

- Units are selected randomly from the sampling frame by a random number generator
- Sample results do not change haphazardly from sample to sample and variability is due to chance

Advantages	Disadvantages
- Sample tends to be a good representation of the population	- Subject to non-response (choose to opt out of the study) - Possible limited access of information as the selected individuals may be located in different geographical location

#### 2. Systematic sampling

- A method of selecting units from a list through the application of selection interval  $K$ , so that every  $K$ th unit on the list, following a random start, is included in the sample

Advantages	Disadvantages
- More straightforward and simpler selection process than the simple random sampling - Do not need to know the exact population size at the planning stage	- May not be representative of the population if the sampling list is non-random

#### 3. Stratified sampling

- Population broke down into strata in which each stratum are similar in nature but size may vary across strata
- Simple random sample is then employed from every stratum

Advantages	Disadvantages
- Able to get a representative sample from every stratum	- Quite complicated and time-consuming  - Need information about sampling frame and stratum, which can be hard to define

#### 4. Cluster sampling

- Breaking down the population into clusters then randomly sample a fixed number of clusters
- Proceed to include all observations from the selected cluster

Advantages	Disadvantages
- Less tedious, costly, and time-consuming as opposed to other sampling methods (eg. stratified sampling)	- High variability due to dissimilar clusters or small numbers of clusters

### Sampling methods - Non-probability sampling

#### 1. Convenience sampling

- Researchers use the subjects most easily available to participate in the research study
- **Selection bias** occurs as certain members of the population may not be included in convenience sampling
- **Non-response bias** occurs as individuals asked to participate in the study may opt out of the study due to inconvenient faced

#### 2. Volunteer sampling

- The researcher actively seeks volunteers to participate in the study
- **Selection bias** might occur as the researcher could pick volunteers more likely to respond in a certain desirable way
- Volunteers sought by the researcher may be unwilling to participate in the study due to inconvenience hence causing **non-response bias**

### Criteria for generalizability:

- Good sampling frame
- Probability-based sampling employed
- Large sample size considered
- There is minimum non-response

## Types of variables

### 1. Dependent vs Independent variable

Independent variable	Dependent variable
A variable that may be subject to manipulation (either deliberately or spontaneously) in a study	A variable which is hypothesized to change depending on how the independent variable is manipulated in the study
Eg. Do NUS students who take notes using pen and paper score better in GEA1000 than those who use laptops? <ul style="list-style-type: none"><li>- Independent variable: Method of note-taking for GEA1000</li><li>- Dependent variable: GEA1000 grades</li></ul>	

### 2. Categorical vs Numerical variable:

Categorical		Numerical	
<ul style="list-style-type: none"><li>- Take category or label values (each observation placed into only one label)</li><li>- Eg. Smoking status</li></ul>		<ul style="list-style-type: none"><li>- Take numerical values for which arithmetic operations such as adding and averaging make sense</li><li>- Eg. Age and mass</li></ul>	
Ordinal	Norminal	Discrete	Continuous
<ul style="list-style-type: none"><li>- Categories that come with natural ordering and numbers used to represent the ordering</li><li>- Eg. Happiness index that is rated 0-10 in order of increasing happiness</li></ul>	<ul style="list-style-type: none"><li>- No intrinsic ordering for the variables</li><li>- Eg. Eye color</li></ul>	<ul style="list-style-type: none"><li>- One where possible values of the variable from a set of numbers with 'gaps'</li><li>- Eg. Population count</li></ul>	<ul style="list-style-type: none"><li>- One that can take on all possible numerical values in a given range or interval</li><li>- Eg. Time, length</li></ul>

## Measures of central tendencies

### 1. Mean (denoted by $\bar{x}$ )

- Given by the formula  $\bar{x} = \frac{X_1 + X_2 + \dots + X_n}{n}$ , where n is the number of data points
- Properties of mean include:

- Adding a constant value to all the data points changes the mean by that constant value
- Multiplying a constant number  $c$  to all the data points will result in the mean also being multiplied by  $c$
- Limitations of mean include
  - Does not tell us about the distribution over the total  $n$
  - Does not tell us about the frequency of occurrence of the values of the numerical variables (refer to pg 23 chapter 1 part 2)
- Mean serves as a good metric when comparing groups of unequal sizes (remember to take the weighted average of sizes of subgroups are different)

## 2. Median

- The middle value of the variable after arranging the values of the data set in ascending/descending order
- Properties of median include:
  - Adding a constant value to all the data points changes the median by that constant value
  - Multiplying all the data points by a constant value,  $c$ , results in the median being multiplied by  $c$
- Limitations of median include:
  - Median alone does not tell us about the total value, the frequency of occurrence or the distribution of data points of the numerical data (similar to mean)
  - Knowing the median of subgroups does not tell us anything about the overall median apart from the fact that it must be between the medians of the subgroups (unlike mean which a weighted average can be taken)

Note: Median is preferably used over mean when the distribution of points is not symmetrical

## 3. Mode

- The value of the variable that appears the most frequently
- Mode can be taken on both numerical and categorical values
- Not very useful when values are unique

## Measure of dispersion (spread) :

### 1. Standard deviation and sample variance

- Used to quantify the 'spread' of data about the mean
- The formula is as such :

$$\text{Sample Variance} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} \quad s_x = \sqrt{\text{Variance}}$$

↓  
Notation for standard deviation of  $x$

- Properties of the standard deviation are as such
  - Always non-negative with the same units as the numerical variable

- Adding a constant value,  $c$ , to all the data points does not change the standard deviation
- Multiplying all the data points by a constant value  $c$  results in the standard deviation being multiplied by the absolute value of  $c$
- When comparing spread relative to the mean, we should utilize the coefficient of variation which is standard deviation of variable/mean of variable

## 2. Interquartile range

- The difference between the third and first quartile
- A small IQR values means that the middle 50% of the data values have a narrow spread and vice versa
- Properties of ICR include:
  - It is always non-negative
  - Adding a constant value to all the data points does not change the ICR
  - Multiplying all the data points by a constant value  $c$  results in the IQR being multiplied by the absolute value of  $c$

## Types of study designs

### Experimental study

Features	<ul style="list-style-type: none"> <li>- Intentionally manipulated one variable in an attempt to cause an effect on another variable</li> <li>- Primary goal is to provide evidence for a cause-and-effect relationship between 2 variables</li> <li>- Subjects are placed in 'treatment' and 'control' groups (in which the 'control' group provides a baseline for comparison)</li> </ul>
The use of random assignment <ul style="list-style-type: none"> <li>- Which is an impartial procedure that uses chance</li> </ul>	<ul style="list-style-type: none"> <li>- Used in order to account for the effects from all the other variables on the study</li> <li>- Use random draw without replacement in which all the chosen subjects belong to the treatment group while the remaining subjects not chosen belong to the control group</li> <li>- If the number of subjects is large, the treatment and control groups will tend to be similar in aspect</li> <li>- The treatment and control group can have different sizes as long as the size of the groups are quite large (then a randomised assignment produces 2 similar groups)</li> </ul>
Blinding	<ul style="list-style-type: none"> <li>- Used to prevent bias as blinded subjects do not know whether they are in the treatment or control group</li> <li>- A placebo very similar to the treatment can be chosen to help make the blinding effect and this is done to prevent the subject's own beliefs about the treatment from affecting the results</li> <li>- Hence treatment and control group would respond the</li> </ul>

	<p>same way to the idea of treatment</p> <ul style="list-style-type: none"> <li>- Blinding of assessors can also take place to prevent bias due to certain pre-conceived notions about the treatment/control group</li> </ul> <p>Note: Double-blinding us when both subjects and assessors are blinded</p>
--	--

### Observational study

Features	<ul style="list-style-type: none"> <li>- The observing of individuals and measuring variables of interest</li> <li>- Researchers do not attempt to directly manipulate one variable to cause an effect in another variable</li> <li>- Does not provide evidence of a cause and effect relationship</li> <li>- Used when a direct investigation is difficult and unethical</li> <li>- Exposure and non-exposure group used to denote the treatment and control groups respectively</li> </ul>
----------	--

Some key differences are:

Experimental	Observational
<ul style="list-style-type: none"> <li>- Assigned by researcher</li> <li>- Can provide evidence of a cause and effect relationship</li> </ul>	<ul style="list-style-type: none"> <li>- Assigned by subjects themselves</li> <li>- Can provide evidence of 'association'</li> </ul>

## Chapter 2

Referring to this table:

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

Marginal rate:

- $\text{Rate}(Y) = 350/1050 = 33.33\%$
- $\text{Rate}(\text{success}) = 831/1050 = 79.1\%$

Conditional rate:

- $\text{Rate}(\text{success} | X) = 542/700 = 77.4\%$

Joint rate:

- This is not a conditional rate
- $\text{Rate}(Y \text{ and failure}) = 61/1050 = 5.81\%$

Association (due to it being an observational study)

Association absent	Association present
$\text{rate}(A B) = \text{rate}(A NB)$ <ul style="list-style-type: none"> <li>- Rate of A is not affected by the presence or absence of B hence A and B are not associated</li> </ul>	<ol style="list-style-type: none"> <li>1. <math>\text{rate}(A B) &gt; \text{rate}(A NB)</math> <ul style="list-style-type: none"> <li>- Presence of A when B is present is stronger than when B is absent hence there is a positive association between A and B</li> </ul> </li> <li>2. <math>\text{rate}(A B) &lt; \text{rate}(A NB)</math> <ul style="list-style-type: none"> <li>- Presence of A when B is present is weaker than when B is absent hence there is a negative association between A and B</li> </ul> </li> </ol>

Rules of rate:

1. Symmetry rule

- $\text{rate}(A|B) > \text{rate}(A|NB) \Leftrightarrow \text{rate}(B|A) > \text{rate}(B|NA)$
- $\text{rate}(A|B) < \text{rate}(A|NB) \Leftrightarrow \text{rate}(B|A) < \text{rate}(B|NA)$
- $\text{rate}(A|B) = \text{rate}(A|NB) \Leftrightarrow \text{rate}(B|A) = \text{rate}(B|NA)$

2. Basic rule of rate: the overall  $\text{rate}(A)$  will always lie between  $\text{rate}(A|B)$  and  $\text{rate}(A|NB)$

3. The closer  $\text{rate}(B)$  is to 100%, the closer  $\text{rate}(A)$  is to  $\text{rate}(A|B)$

4. If  $\text{rate}(B) = 50\%$ , then  $\text{rate}(A) = \{\text{rate}(A|B) + \text{rate}(A|NB)\}/2$

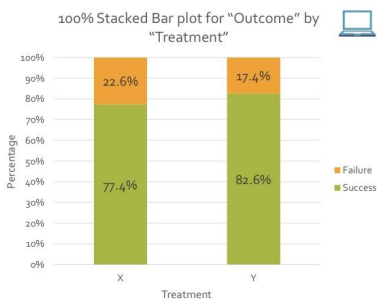
5. If  $\text{rate}(A|B) = \text{rate}(A|NB)$ , then  $\text{rate}(A) = \text{rate}(A|B) = \text{rate}(A|NB)$

Simpson's paradox

- A phenomenon in which a trend appears in several groups of data but disappears or reverses when the groups are combined

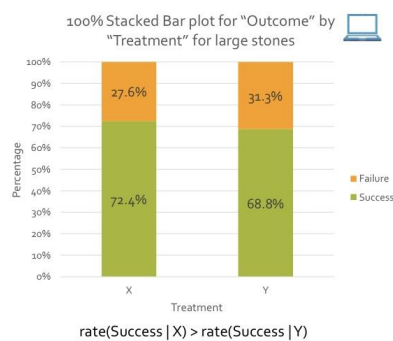


- When the majority of the individual subgroup rates are opposite from the overall association



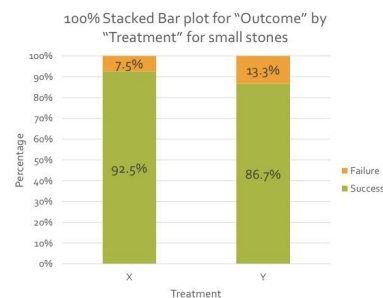
**Y** Overall, Treatment Y is better

All stones	Success	Failure	Total
X	542	158	700
Y	289	61	350
Total	831	219	1050



**X** Across large stones, Treatment X is better

Large stones	Success	Failure	Total
X	381	145	526
Y	55	25	80
Total	436	170	606



**X** Across small stones, Treatment X is better

Small stones	Success	Failure	Total
X	161	13	174
Y	234	36	270
Total	395	49	444

This shows that treatment Y was positively associated with success overall but individually across large and small stones, treatment X was positively associated with success

Confounders:

- Known as a third variable that is associated to both the independent and dependent variable whose relationship we are investigating
- Simpson's paradox implies the presence of confounders but the converse is NOT TRUE
- **Slicing** is a method to control confounders (randomisation not always possible in studies)

## Chapter 3

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

- First step in EDA is to create a graph of the distribution of a variable, most commonly a histogram or boxplot (univariate EDA)

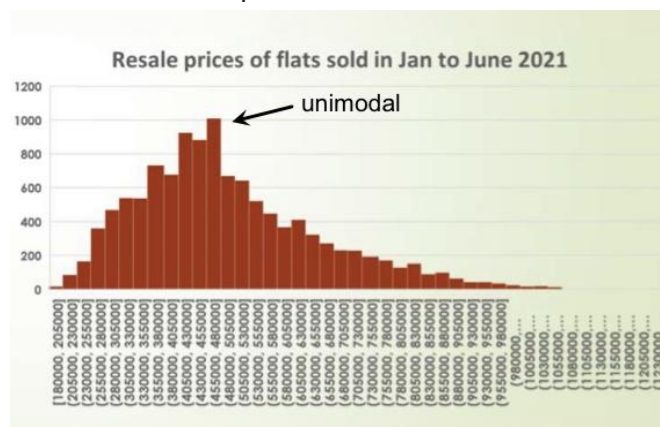
### Histograms:

- Presents a graphical display of a distribution + quick/easy to grasp
- Useful for large data sets as it is impractical to show the value of each observation

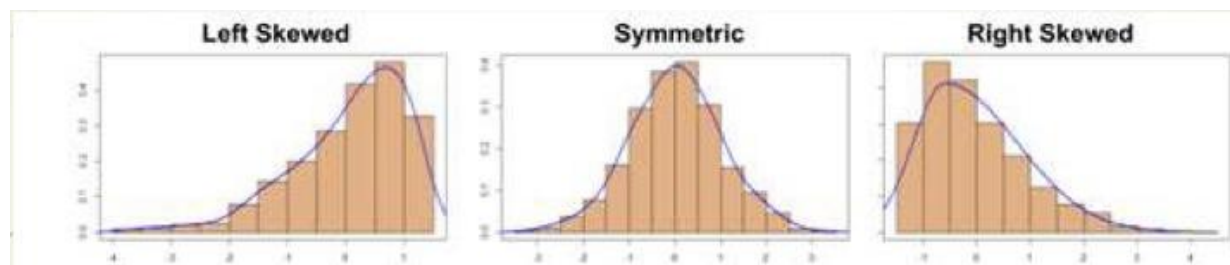
When describing a histogram, we need to note:

#### 1. Shape of a distribution : peaks and skewness

- There is a peak in the interval in a unimodal distribution, meaning it has one distinct peak

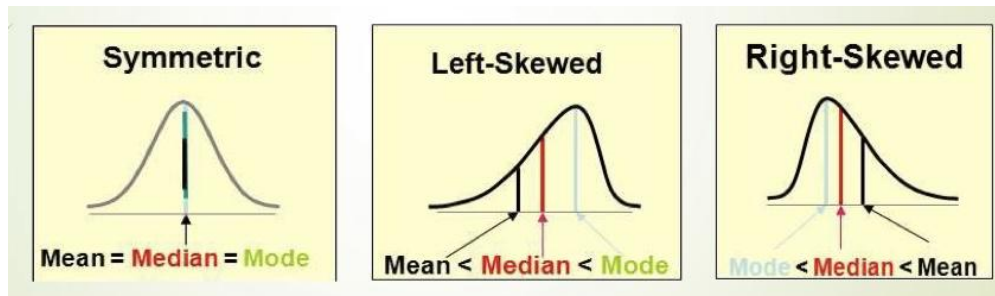


- Next, in a symmetrical distribution, its left and right halves are mirror images of each other and the peak is in the middle. (see : normal distribution curve)
- If left skewed, its peak is shifted towards the right and if it is right skewed, its peak is shifted towards the left



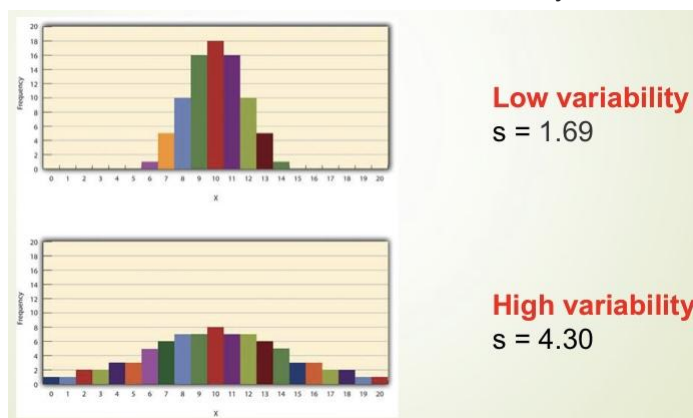
#### 2. Centre of a distribution : mean, median and mode

- In a symmetrical distribution, the mean, the median and the mode will be very close to each other at the peak of the distribution
- In a left skewed distribution,  $\text{mean} < \text{median} < \text{mode}$
- In a right skewed distribution,  $\text{mean} > \text{median} > \text{mode}$



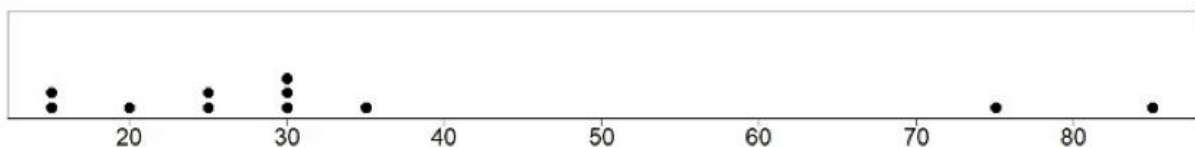
### 3. Spread of a distribution : range and standard deviation

- The higher the variability, the wider the range in which the data is being spread across
- Most common measure of variability is the standard deviation



### 4. Outliers : observations that fall well above or well below the overall bulk of data

- Examining data for outliers can be useful in identifying strong skew in a distribution, identifying possible data collection/data entry errors or for providing interesting insight into the data
- Mean can be pulled so far in the direction of skew that it may no longer be a good measure of the central tendency of that distribution



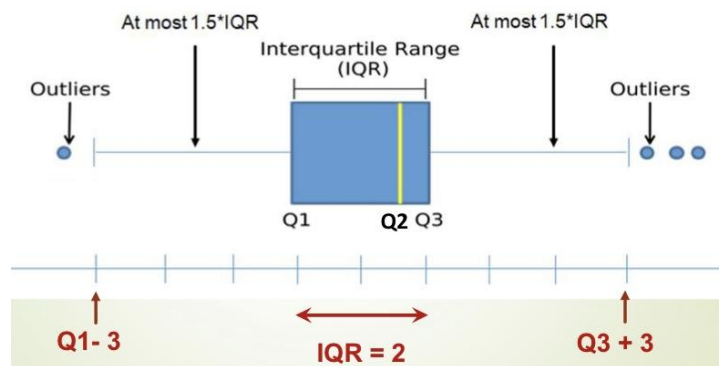
### More about histograms :

- Avoid histograms with large bin widths as it does not give good information about the variability in the distribution
- Avoid histograms with small bin widths as we may have a broken comb look that does not give a sense of the distribution

- A histogram shows the distribution of numerical variable across a number line but a bar graph makes comparison across categories of a variable
- The ordering of bars in a histogram cannot be changed unlike those of a bar graph
- There are no gaps between bars in histograms

### Boxplots:

- We use the minimum, quartile 1 (Q1), median, quartile 3 (Q3) and the maximum to construct a boxplot
- A data point is considered an outlier if its value is greater than  $[Q1 + 1.5] \cdot IQR$  or if its value is less than  $[Q1 - 1.5] \cdot IQR$



### 1. Shape

- Shape can be deduced by comparing the variability in the upper half of the data (max-median) to the variability in the lower half of the data (median-min)
- Skewed right if lower half has less variability than the upper half and vice versa

### 2. Center

- Can deduce the median value at a glance compared to that of a histogram

### 3. Spread

- IQR gives an idea of the spread for the middle 50% of the data set

### Comparisons:

Histogram	Boxplot
<ul style="list-style-type: none"> <li>- Provide a better sense of the shape of distribution of a variable, especially when there are great differences among the frequencies of the data point</li> </ul>	<ul style="list-style-type: none"> <li>- More useful when comparing distributions of different data sets</li> <li>- Identify and exhibit outliers clearly</li> </ul> <p>However, they do not give any information about how many data points we are working with (2 boxplots look the same but correspond to data sets with very different</p>

	numbers of data points)
--	-------------------------

## Bivariate EDA

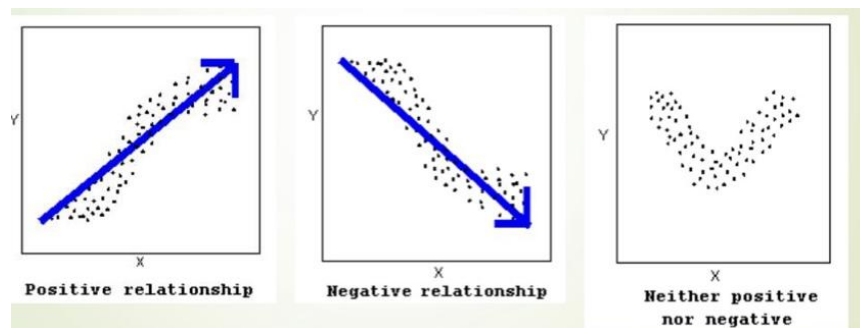
Deterministic relationship	Statistical/non-deterministic relationship (association)
<ul style="list-style-type: none"> <li>- One numerical value or quantitative variable can be used to determine another</li> </ul>	<ul style="list-style-type: none"> <li>- Existence of variability makes it impossible to find a unique value of one variable corresponding to each value of the other variable</li> <li>- Average value of one variable is described given the value of the other variable (known as association)</li> </ul>

Tools used to analyse bivariate data include:

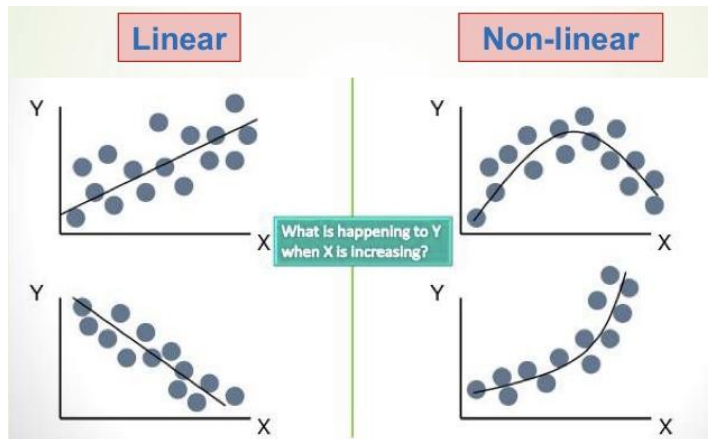
### 1. Scatter plots

- To have an idea of the pattern formed between 2 variables
- In order to interpret scatter plots, we can consider
  - Direction
    - The direction can be positive, negative or neither
    - A positive/increasing relationship means that an increase in one of the variables is associated with an increase in the other
    - A negative/decreasing relationship means that an increase in one of the variables is associated with a decrease in the other

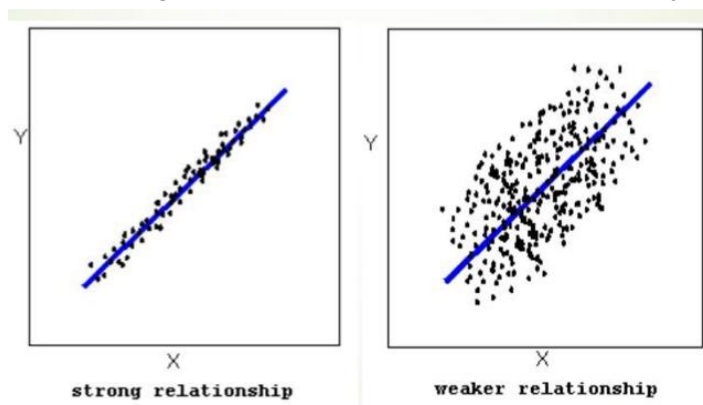
Note: Not all relationships can be classified as either positive or negative



- Form (general shape)
  - Linear or non-linear form (data point scattered about a line or the data points appear scattered about a smooth curve)



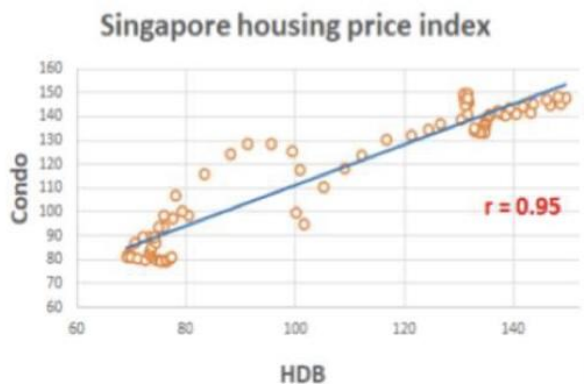

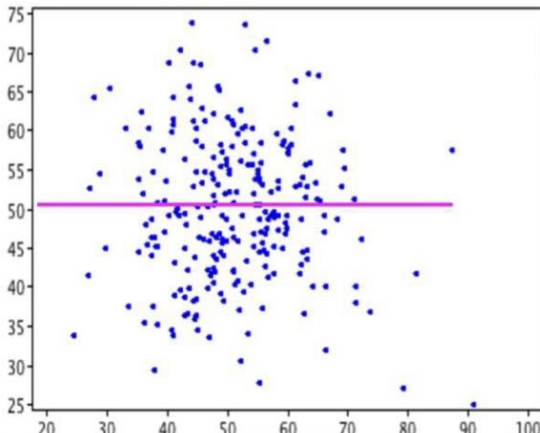
- Strength
  - How closely the data follow the form of the relationship
  - Following the line pattern closely will be seen as a strong relationship while the points being more scattered about the line will imply a weaker relationship



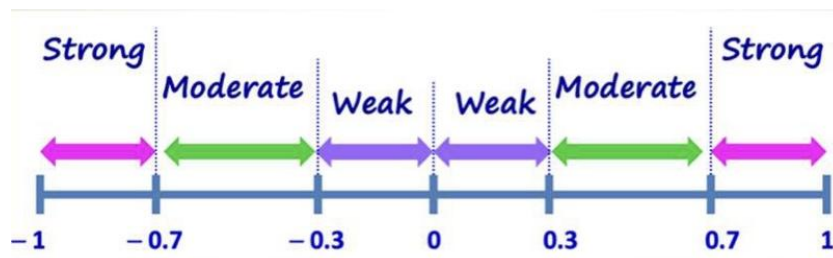
- Outliers
  - Points that deviate from the pattern of the relationship

## 2. Correlation coefficients (a measure of linear association)

- To check if the data are linearly related
- It has a range between -1 and 1
- It summarizes direction and strength of linear association

<p><math>r &gt; 0</math> positive association</p>	<p>Singapore housing price index</p> 
<p><math>r &lt; 0</math> negative association</p>	<p>Gold vs Oil</p> 
<p><math>r = 0</math> no linear association</p>	

Note:



- As the value of  $r$  gets closer to 1 or -1, the data falls more closely to a straight line

In order to calculate the correlation coefficient:

- First, convert each data point into its standard unit
- $SU_X = \frac{X - \text{average}(X)}{s_x}$  and  $SU_Y = \frac{Y - \text{average}(Y)}{s_y}$  where  $s_x$  is the standard deviation of  $X$  and  $s_y$  is the standard deviation of  $Y$
- Second,  $r$  value is just the average of the product of  $X$  and  $Y$  in standard units.

Also ,

- $r$  is not affected by the interchanging of two variables, the adding of a number to all values of a variable and the multiplying a positive number to all values of a variable
- Correlation does not imply causation ( $r$  value close to 1 to -1 implies strong statistical relationship NOT a causal relationship)
- $r$  value might still be present for non-linear associations between two variables but  $r$  only supposed to measure linear association hence LOOK at the scatter plot and not just the  $r$  value
- Outliers can increase/decrease the strength of the correlation coefficient (refer to pg79,80 of lecture 3)

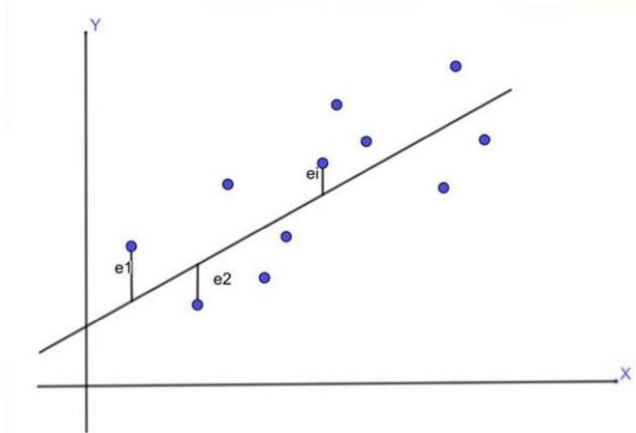
### 3. Regression analysis

- To fit a line or curve to a data set and do predictions using the data
- We model the relationship by a straight line  $Y = mX + b$ , where the constant  $b$  refers to the  $y$  intercept and  $m$  refers to the gradient
- The slope of the regression line and the correlation coefficient is related by  $m = \frac{s_y}{s_x} r$  where  $s_y$  is the standard deviation for  $y$  and  $s_x$  is the standard deviation for  $x$

To find the regression line:

- Define the  $i$ -th residual of the observation : where  $e_i$  = difference between the observed outcome and predicted outcome in which we want to minimize  $e_1^2 + \dots + e_n^2$  where  $n$  = the number of data points
- The regression line is then obtained by minimizing the sum of the squared errors





The value of  $y$  we obtain from the equation is NOT the exact value and the equation also CANNOT be used to predict the 'x variable'. This is because

- The equation for predicting the  $y$  variable based on the  $x$  variable is obtained by minimizing the sum of squared errors for  $y$  variable
- Also, prediction of  $y$  variable beyond the observed range of the  $x$  variable is dangerous as the regression line may change

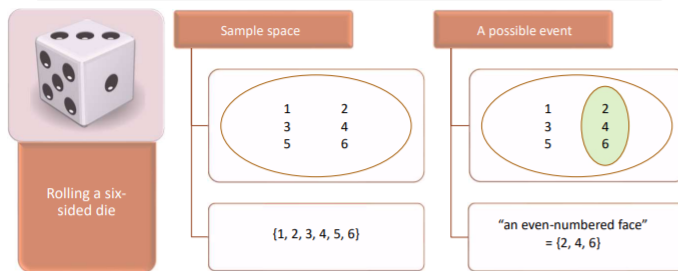
## Chapter 4

### Probability - a prelude to statistical inference

#### Definitions

Sample Space	A collection of all outcomes of a probability experiment
Event	Subcollection of the sample space
Sample statistic	<p>Refers to the use of a sample to draw a conclusion about the population</p> <p>Since a simple random sampling with 100% response rate is used the expression is as such</p> <ul style="list-style-type: none"> <li>- Sample statistic = population parameter + random error</li> <li>- There will be no selection bias or non-response bias unlike a sample estimate of the population parameter</li> </ul>

## Example: Die-rolling



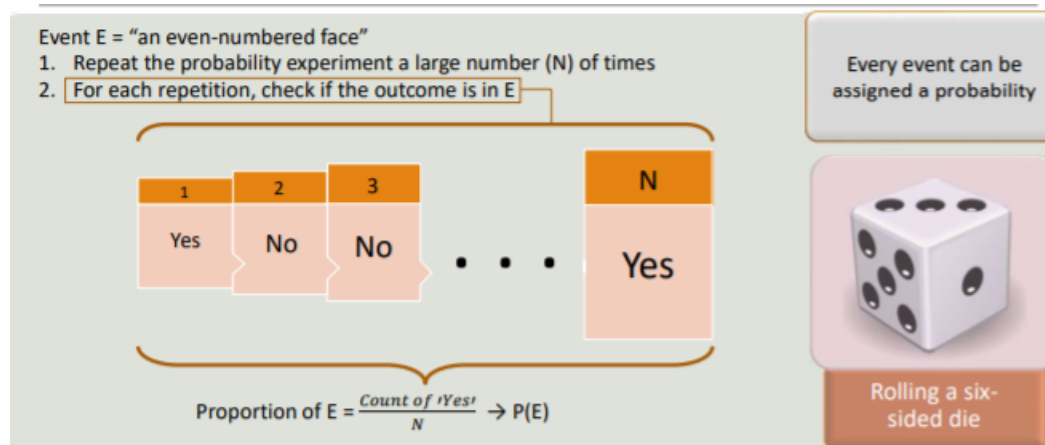
A probability experiment must:

- Be repeatable ("as many times as you want")
- Give rise to a precise set of outcomes

If E is an event that has been assigned a probability:

- $P(E)$ , read "the probability of E", stands for the probability assigned to E
- Numerical value between 0 and 1 inclusive

## Finite Sample Spaces



Note:

Proportion of E estimates the true  $P(E)$ .

All these proportions are estimates of the true probability of E, and such estimates get more **accurate** as N gets larger.

Rules of Probabilities:

1.  $0 \leq P(E) \leq 1$  for each event E
2.  $P(S) = 1$  if S is the entire sample space

3. If E and F are non-overlapping (mutually exclusive) events, then  $P(E \cup F) = P(E) + P(F)$

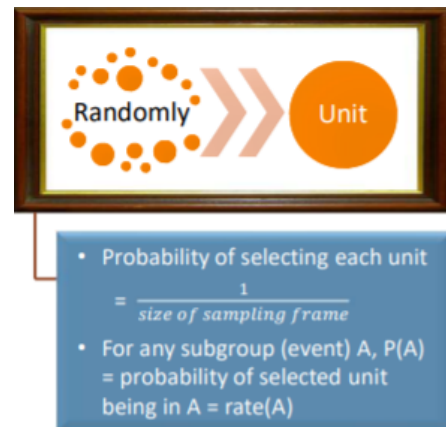
For finite spaces, it is enough to assign probabilities to outcomes so that they add up to 1.

#### Uniform Probabilities and Rates:

Uniform probability assigns equal probability to every outcome, so that each outcome has the probability one divided by the size of the sample space.

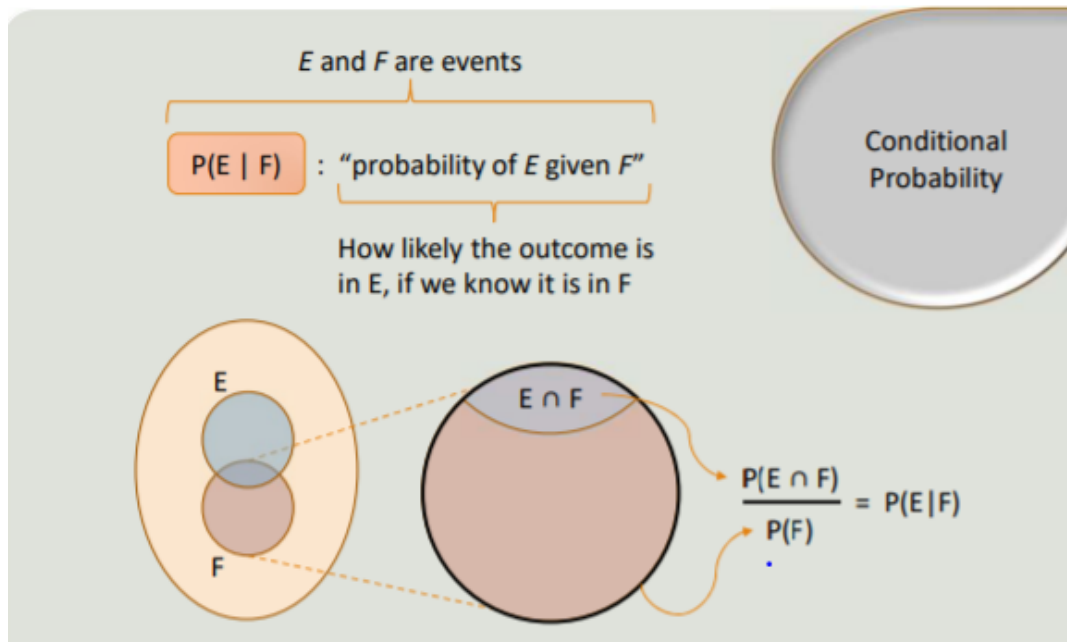
Every outcome has the same probability

$$= \frac{1}{\text{size of sample space}}$$



This is also used in simple random sampling, when we randomly select a unit from a sampling frame, we are conducting a probability experiment, and the sample space of this probability experiment is exactly our sampling frame.

#### Conditional Probabilities:



Note:

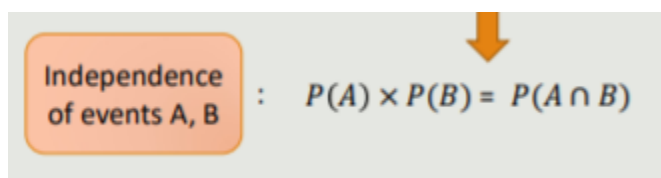
By convention, if  $P(F) = 0$ , then  $P(E | F) = 0$  for any  $E$ .

$P(E | F) = \text{rate}(E | F)$

Independence:

Defining Independence of two events  $A$  and  $B$ :

- probability of  $A$  equals the probability of  $A$  given  $B$ .

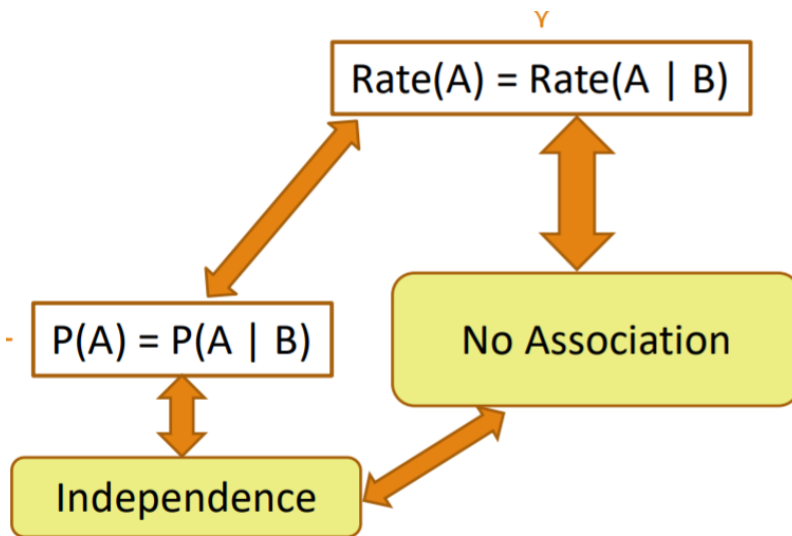


Note:

Order of events does not matter when talking about independence.

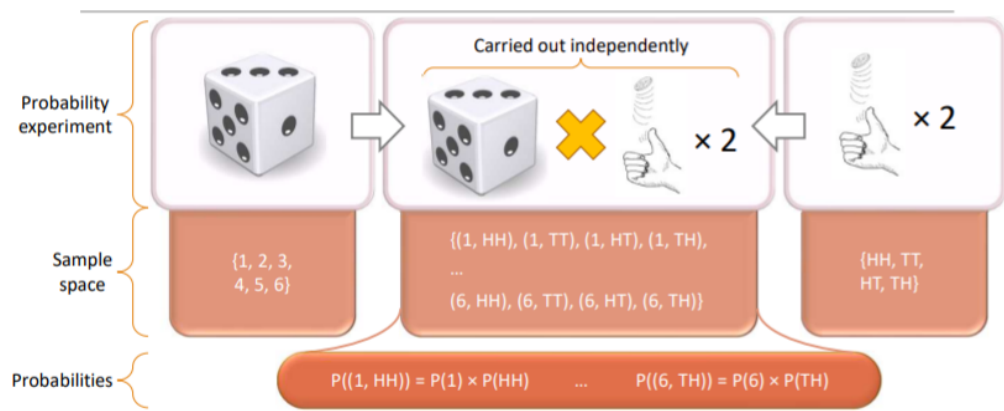
Independence and Association:

By the basic rule of rates, the rate of  $A$  being equal to the rate of  $A$  among  $B$  precisely indicates the lack of association between the two independent variables.



### Independent Probability Experiments

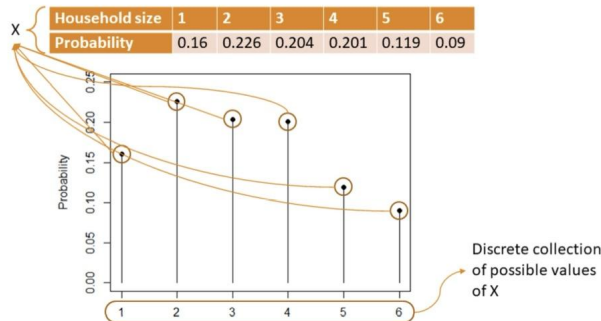
- Two probability experiments that are independent
- The combined probability experiment has sample space with all pairings of possible outcomes of the two components



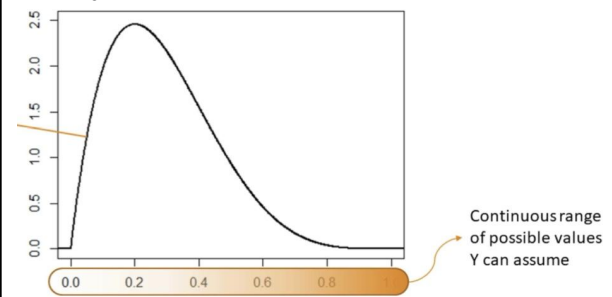
### Random variables

Discrete random variables	Continuous random variable
<ul style="list-style-type: none"> <li>- Points in the plot separated by gaps hence the x values are discrete</li> <li>- In accordance with the rules of probabilities such that the probabilities of the points add up to 1</li> </ul>	<ul style="list-style-type: none"> <li>- Any continuous random variable Y can be visualised with a density curve on the standard x and y axis in which a curve can be viewed as 'continuous series of points'</li> <li>- The x value of a highest point of the curve</li> </ul>

- the mode of a discrete random variable is the x values of the highest point
- Eg.  $P(X \geq 5) = P(5) + P(6)$



- as a mode
- $P(0.3 \leq Y \leq 0.5)$  = shaded area under the density curve Y in the interval 0.3 to 0.5



## Normal distributions

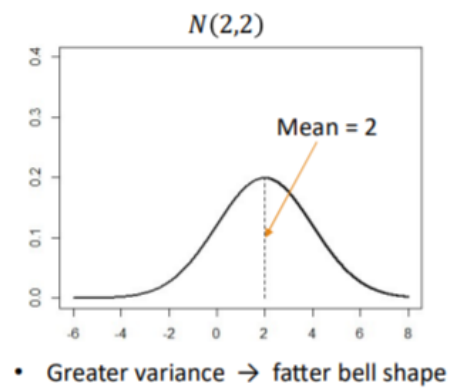
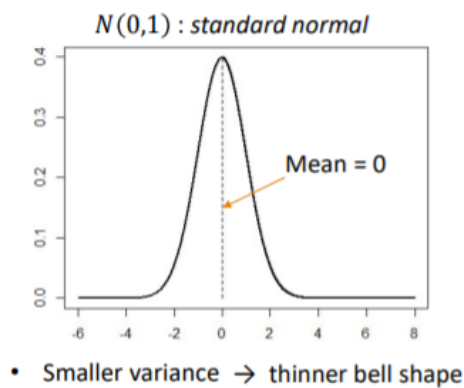
$N(x, y)$ : normal distribution with mean x and variance y

Normal Distributions can only differ by their means or variances

Properties:

- Bell-shaped curve
- Peak of curve occurs at the mean
- Curve is symmetrical about the mean

Note: Mean = Mode = Median of any normal distributions

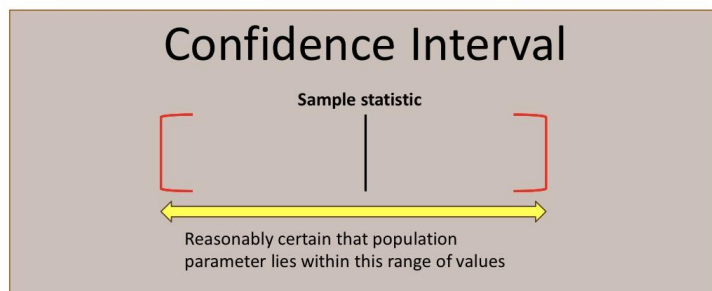
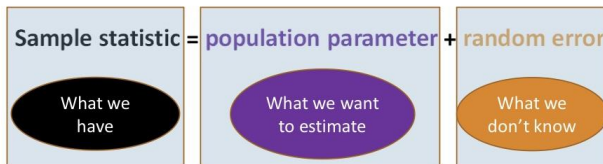


Area under curve kept constant  $\rightarrow$  a fatter curve compensates by being shorter

Note: Area under curve always = 1.

## Statistical inference - Confidence Intervals

- Used to make statements on how confident we are of the sample statistic being used to estimate the population parameter
- A repeatable means of generating an interval based on a simple random sample drawn from the population
- Provides a range of values that we are reasonably certain that the population parameter lies in



### C.I. for Population Proportion

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

where  $p^*$  is the sample proportion,  $z^*$  is the value from the standard normal distribution and  $n$  is the sample size

$$\bullet \text{Margin of error: } z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

• For 90% C.I.,  $z^*$  is 1.645

• For 95% C.I.,  $z^*$  is 1.96

• For 99% C.I.,  $z^*$  is 2.58

### Interpreting C.I.

- We present our C.I. in 2 parts: our confidence level and the interval
- Example:  
xx% C.I.:  $p^* \pm (\text{margin of error})$
- We are xx% confident that the population parameter lies in the confidence interval

Confidence Level -> xx%

Confidence Interval ->  $p^* \pm (\text{margin of error})$

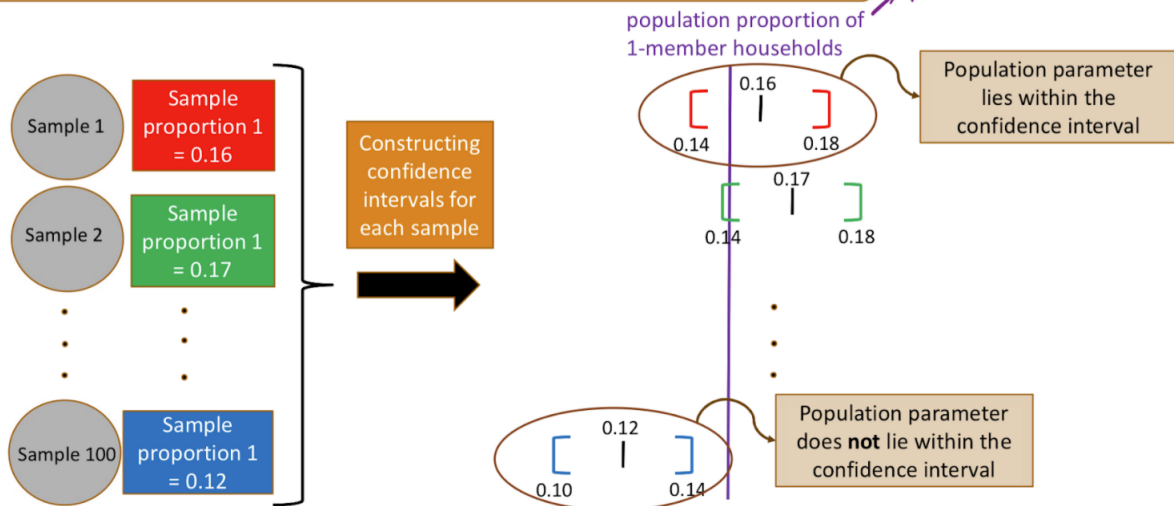
Some properties of confidence interval include:

- The larger the sample size, the smaller the margin of error, the smaller the confidence interval
- Sample statistic will **always** be within your confidence interval

## Sample 100 - 95% CI: $0.12 \pm 0.02$

"95% confident that the population parameter lies within the confidence interval"

[population parameter]



Understanding "Confidence": [Assume 95% confidence in this case]

- Does not mean that there is a 95% chance that our population parameter will fall in our confidence interval
- Instead, it means that if we collect **many random samples** and **construct a confidence interval for each of them**, about 95% of them would contain the population parameter

Eg: Referring to the above case, if you take 100 simple random samples from the population and calculate their confidence intervals, 95 confidence intervals (bracketed ranges of the samples) would contain the population parameter (population proportion of 1-member households - denoted by purple line).

## Statistical inference - Hypothesis testing

Step 1 : Identifying the question and state both the null and alternative hypotheses

- Produce 2 hypotheses namely the null and alternative hypotheses
- Null hypothesis takes a stance of no effect (assume 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 put against the null hypothesis
- Both hypotheses must be mutually exclusive



Step 2 : Collect the relevant data, as well as deciding 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
- In a coin tossing example, the random variable is the number of heads out of 8 independent coin tosses, assuming coin is fair

Step 3 : Determining the level of significance (between 1 and 0) as well as 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
- 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 seen as the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true

Example : Say we observe 7 heads out of 8 toss

$$\begin{aligned}\square \text{ p-value} &= P(\text{obtaining a result at least as extreme as observed} \mid \text{null hypothesis is true}) \\ &= P(7 \text{ heads out of 8 heads is true} \mid \text{null hypothesis is true}) + \\ &\quad P(8 \text{ heads out of 8 heads is true} \mid \text{null hypothesis is true}) \\ &= 8 \left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^8 = 9\left(\frac{1}{2}\right)^8 = 0.035156.\end{aligned}$$

Step 4 : Make conclusion about the null hypothesis, whether we reject it or not

- Reject the null hypothesis in favour of the alternative , if p value < significance level
- Do not reject the null hypothesis if p value  $\geq$  significance level and the test result is inconclusive
- If p value is not lower than the level of significance , we cannot reject the null hypothesis which means we don't know if observation is due to chance or not
- Not rejecting the null hypothesis does not make it true
- We cannot reject the alternative hypothesis

One sample t-test:

- Require for the population distribution to be approximated normal, if n, the sample size is smaller than 30
- The data used should be produced randomly

Example:

☐ Null hypothesis,  $H_0: \mu = 600$  thousand.

☐ Alternative hypothesis,  $H_1: \mu > 600$  thousand.

☐ Using Rcmdr, we see that p-value = 0.0004832 < 0.05.



☐ At the 5% level of significance, we have sufficient evidence to reject the null hypothesis and accept the alternative hypothesis.

☐ Meaning to say, we accept researcher B's claim that the average price of HDB flats in Singapore in the year 2020 is higher than 600 thousand dollars.

Chi-squared test:

- Use to check whether 2 categorical variables are significantly associated
- The data must be counts for the categories of a categorical variable
- We would like the cases within each group to be selected randomly

Example:

☐ Null hypothesis,  $H_0$ : Surgery success is not associated with treatment type among all kidney patients in the hospital.

☐ Alternative hypothesis,  $H_1$ : Surgery success is associated with treatment type among all kidney patients in the hospital.

☐ We use chi-squared test on this sample to determine if treatment and success are significantly associated at the population level.

☐ The p-value is 0.07991. We do not have enough evidence to reject the null hypothesis at the 5% level of significance.

☐ We cannot conclude that treatment type is significantly associated with outcome at the population level.

### Comparison between the different hypothesis test

One sample t-test	Chi-squared test
<ul style="list-style-type: none"><li>- Mainly used when testing for significant difference between sample mean and a known/hypothesized mean</li><li>- Population distribution should be approximately normal if n, the sample size is smaller than 30</li><li>- Data used is acquired randomly</li></ul>	<ul style="list-style-type: none"><li>- Mainly used when testing for significant association between 2 categorical variables</li><li>- The data is given is the count for the categories of a categorical variable</li><li>- Data used is acquired randomly</li></ul>