

Contents

Chapter 1 - Quantitative Reasoning 2

Chapter 2 – Categorical Data Analysis 13

Chapter 3 – Dealing with Numerical data 16

Chapter 4 - Statistical Inference..... 32

Chapter 1 - Quantitative Reasoning

Population – The population of the entire group that we want to know something about.

Parameter – A numerical fact about a Population

Research Question

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	Are student athletes more likely than non-athletes to do final year projects?
Investigate a relationship	Is there a relationship between the average number of hours students spend each week on Facebook and their GPA?

Population of Interest – A group in which researcher has interest in drawing

Example: “population of Asia”, “population of Singapore”, “population of Ang Mo Kio”

Sample – A proportion of the population selected in the study

Sampling Frame – “Source Material” from which sample is drawn from.

- May not cover the population of interest, or may contain units that are not in the population of interest

Sampling frame generalizability criteria: Sampling frame should be equal to, or larger than the target population.

- Whether the results can be generalized to some larger population will depend on the sampling method used

Census – An attempt to reach out to entire population of interest. A census with 100% response rate and no response-bias is the ideal method of exactly determining the population parameter.

Sample – A proportion of the population selected. Sampling is preferred over collection of entire population because of cost and speed. In most cases, it is impossible to determine the parameter exactly, but it can be estimated using a sample.

Factors needed for a good estimate

- 1) **Sampling frame must contain the population of interest***
- 2) **Probability sampling*** (All non-probability sampling results in selection bias!)
- 3) Must be large enough
- 4) **High response rate**

***If these two factors are not enforced, we encounter selection bias**

Estimate = parameter + bias + random error

Bias – Influenced by 1, 2, 4. Aim for minimal selection bias and non-response

Random Error – Influenced by 3. Larger the sample, smaller this error.

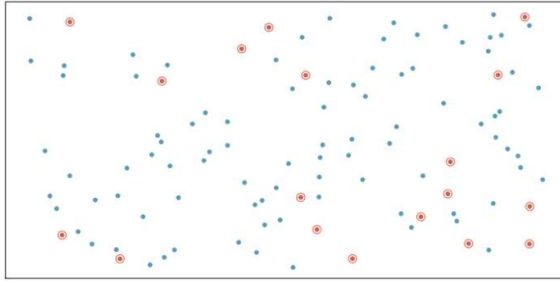
Bias

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
<ul style="list-style-type: none">• Imperfect Sampling Frame• Non-Probability Sampling	<ul style="list-style-type: none">• Disinterested• Inconvenient• Unwilling to disclose sensitive information

Sampling Methods

Probability Sampling – The selection process is via a known randomized mechanism. Probability may not be the same throughout all units of population. Eliminates biases associated with selection. Every unit in the population has a non-zero and known probability of being selected.

Simple Random Sampling



- Units are randomly selected from the sampling frame (w/ or w/o replacement)
- Mechanism: Random number generator
- Variability is due to chance.

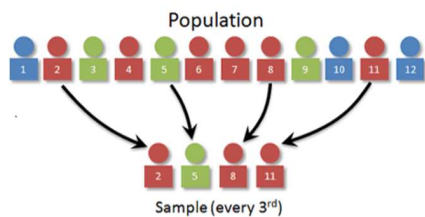
Advantages: Sample tends to be a good representation of population

Disadvantages: Subject to non-response; accessibility of information

Sampling with Replacement

- Same chance of selecting for every unit
- Possibility of units being chosen more than once
- ^ To overcome that with larger population size, smaller samples in proportion to population size

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

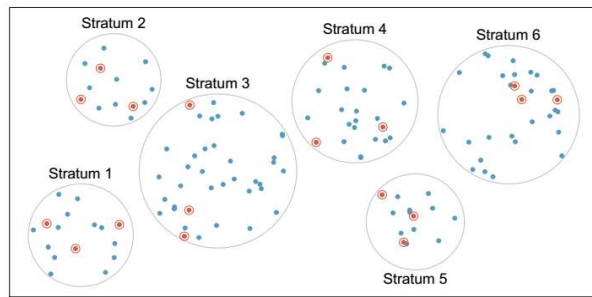


Advantages: Simpler selection process than simple random sampling

Disadvantages: May not be representative of population if list is non-random (potentially under-representing the population)

Another advantage is that for systematic sampling, we may not need to know the exact population size at the planning stage. If we have a rough estimate of the number of dormitory residents in the population, our systematic sample can still produce the same results as a simple random sample.

Stratified Sampling

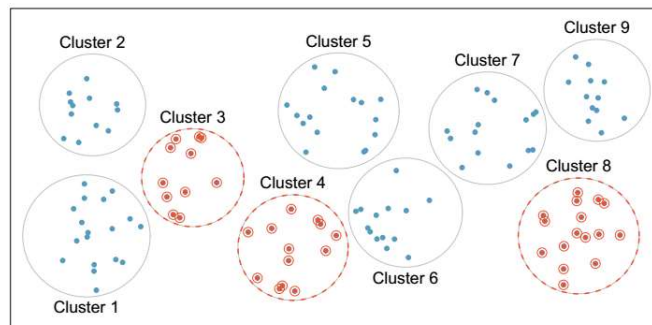


- Breaking down the population into strata
- Each stratum is similar in nature, but size may vary across strata
- Simple random sample from every stratum
- Example: Sample Count (General Election)
- Good to use if one also wants estimates within subgroups
- Estimate of parameter is done via taking weighted average of subgroup estimates

Advantages: Able to get a representative from every stratum

Disadvantages: Need information about sampling frame and stratum (Complicated and time-consuming)

Cluster Sampling



- Breaking down the population into clusters
- Randomly sample a fixed number of clusters
- Include all observations from selected cluster
- Example: Mental wellness survey in schools

Advantages: Less tedious, time-consuming and costly

Disadvantages: High variability due to dissimilar clusters or small number of clusters (Requires larger sample size in order to achieve low margin of error)

Non-Probability Sampling – The selection of individuals/unit were not done by randomization, but by human discretion

Convenience Sampling

Non-probability sampling method in which the researcher uses the subjects that are most easily available to participate in the research study. Individuals may not respond, leading to non-response bias. There can also be selection bias as some groups of people might be left out.

Volunteer Sampling

It is a non-probability sampling method in which the researcher actively seek volunteers to participate in the study.

Conclusion

- Decide on the sampling frame
- Decide whether to employ probability sampling in the chosen sampling frame based on feasibility
- Choose people/unit from the sampling frame
- Removed unwanted units from the sampling frame or the sample, if the sampling frame includes people who do not belong to the target population

Generalizability Criteria

- Good sampling frame (equal or larger than target population)
- Probability-based sampling (reduces selection bias)
- Large sample size (reduces variability of data)
- Minimum non-response

Variables

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

Independent variable – Subject to manipulation (either deliberately or spontaneously) in a study

Dependent variable – Hypothesized to change depending on how the independent variable is manipulated in a study

Example: 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

Categorical variables – Each observation can be placed in only one label. Labels are mutually exclusive

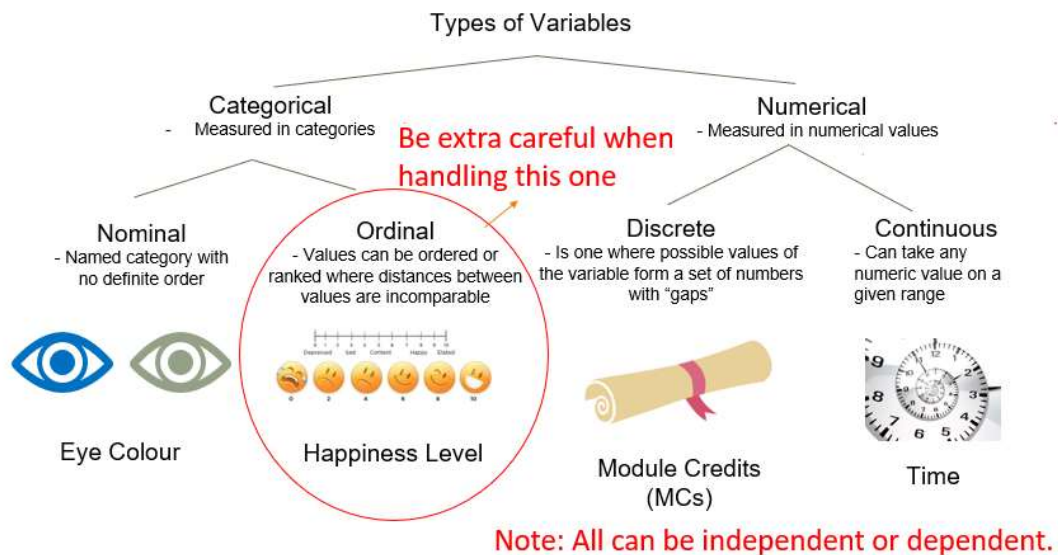
Ordinal – Natural ordering and numbers are often used to represent ordering.

Nominal – No intrinsic ordering. Ex: eye colour

Numerical variables – Numerical values for which arithmetic operations make sense

Discrete – Possible values of the variable form a set of numbers with gaps

Continuous – Can take any numeric value on a given range



Summary Statistics

Micro – Get information on a particular individual(s)

Macro – Get information on groups/population

Central Tendencies – Central or typical value for a probability distribution.

- Mean
- Median
- Mode

Measures of dispersion

- Standard deviation
- Interquartile range

Mean

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \text{ which is also written as } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Properties

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

What the mean can tell us

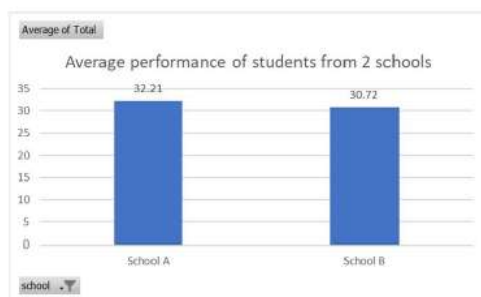
- 1) The mean allows the total to be easily calculated. (Only the mean and number of data points is required)

What the mean can't tell us

- 1) Distribution of data

Overall mean vs means in subgroups

Overall mean is known as taking a weighted average. The overall mean will lie between the largest and smallest subgroup means.



The bar graph shows the average performance of students when categorized by school. The maximum score attainable is 60.

Number of students	
School A	349
School B	46
Total	395

The weighted average of the above is: $\frac{349}{395} * 32.21 + \frac{46}{395} * 30.72 = 32.03$

Sample Variance & Standard Deviation

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

The standard deviation is one way of quantifying the “spread” of the data about the mean. The formula is derived via the variance.

Properties of Standard deviation

- The standard deviation is always non-negative (≥ 0), with same units as the numerical variable.
- Adding a constant value, c (positive or negative) to all the data points does not change the standard deviation.
- Multiplying all the data-points by a constant value c result in the standard deviation being multiplied by $|c|$ where $|c|$ is the absolute value of c .

Coefficient of variation

$\text{coefficient of variation} = \frac{s_x}{\text{mean of } x}$ The coefficient of variation is a way of quantifying the degree of spread, relative to the mean. The formula is (S.D. of variable) / (mean of variable).

Median

The median of a numerical variable in a dataset is the middle value of the variable after arranging the values of the dataset in ascending/descending order. If there are even number of values, the median is the average of the middle 2 values.

Properties of Median

- Adding a constant value (positive or negative) 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 .

What the median can tell us

- 1) Ex. The median score is 30.5. This translates to 50% of students scoring above 30.5, and 50% below 30.5. The mean doesn't give this information.

What the median can't tell us

- 1) Distribution
- 2) Total Value
- 3) Frequency of occurrence

Overall median vs median in subgroups

Median shares same property as means whereby the overall median is between the smallest and largest medians of subgroup. However, unlike the mean, the weighted average of subgroup cannot be used to calculate overall median.

Quartiles and Interquartile Range

The IQR is another way of quantifying the spread of the data.

- The first quartile usually denoted by Q_1 is the 25th percentile of the data-values.
- The third quartile, usually denoted by Q_3 is the 75th percentile of the data-values.
- The interquartile range is the difference between the third quartile and the first quartile.
 $IQR = Q_3 - Q_1$. 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.

Similarities between IQR & S.D

- The IQR & S.D is always non-negative (≥ 0), because Q_3 is at least as large as Q_1 .
- Adding a constant value, c (positive or negative) to all the data points does not change the IQR & S.D.
- Multiplying all the data-points by a constant value c result in the IQR/S.D being multiplied by $|c|$ where $|c|$ is the absolute value of c .
- Data with similar S.D or IQR might not necessary have similar spread patterns

The median is often used in preference to the mean when the distribution of points is not symmetrical.

Mode

Mode of a variable is the value of the variable that appears the most frequently.

- Whilst the mean and median apply strictly for numerical values, the mode can take on both numerical and categorical values.
- The mode is generally interpreted as the peak of a distribution.

Application of Mode

- Real Estate: Real estate agents need the mode of the number of bedrooms per house so they can inform their clients on how many bedrooms they can expect to have in houses located in a particular area.
- HR : Human Resource managers also use the mode of different positions in the company so that they can be aware of the most common position of employees at their company.

Study Designs

There are two main study designs – **Experimental** and **Observational**.

Experiments

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***.

Independent Variable -affect> Dependent Variable

- It is important to have a **treatment group and control group**.
- To establish a cause-and-effect relationship, we want to make sure that the independent variable is the only factor that impacts of the dependent variable.

Random Assignment

To account for the effects from the other variables of concern – Random Assignment. **Random assignment is an impartial procedure that uses chance.** 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 goal of random assignment is to create similar treatment and control group.

Placebo

- 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.

Blinding

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.
- An experiment is called **double-blind if both subjects and assessors are blinded** about the assignment.
- In a **single-blinded experiment, either the participants OR evaluators** are blinded.

In order of effectiveness,

- **No Control**
- **Treatment and Control (w/o Randomization)**
- **Randomised Controlled**
- **Single-Blinded Randomised Controlled**
- **Double blinded randomised Control**

Observational

Observational studies are an alternative approach to experiments especially for scenarios where there are ethical issues preventing the use of experiments.

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. ***It does not provide convincing evidence of a cause-and-effect relationship.*** However, they can still provide evidence of 'association'.

Controlled experiment	Observational study
Independent variable - Mainly consists of treatments and controls	Independent variable - Also known as the exposure variable
Dependent variable - Also known as outcome/response/disease variable	Dependent variable - Also known as outcome/response/disease variable

	Observational Study	Experiment
Assignment	Subjects Self-assigning	Researchers decide
Confounding factors	Likely	Unlikely if random assignment is done
Ethical issues	Unlikely	Possible <i>(if treatment is harmful to some participants)</i>
Possible to show causation	No	Yes in the ideal scenario
Able to show association	Yes	

Chapter 2 – Categorical Data Analysis

PPDAC Cycle

- Problem
- Plan
- Data
- Analysis
- Conclusion

Rates

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

Marginal Rate

Rate(Y) = $350/1050 = 33.3\%$

Rate(Success) = $831/1050 = 79.1\%$

Conditional Rate

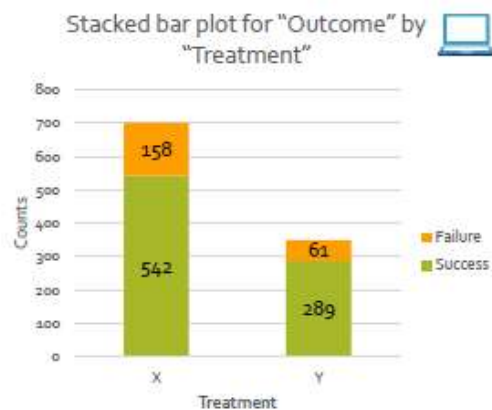
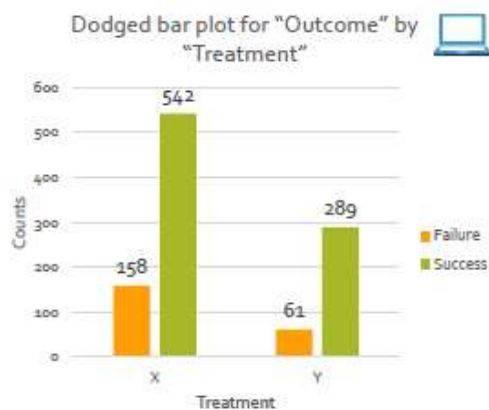
Rate(Success | X) = $542/700 = 77.4\%$

Joint Rate

Rate(Y and Failure) = $61/1050 = 5.81\%$

- Joint rate is **NOT** conditional rate

Bar plots



Association

Association absent

$$\text{Rate}(A|B) = \text{rate}(A|NB)$$

A and B are not associated.

Association present

$$\text{rate}(A|B) \neq \text{rate}(A|NB)$$

$\text{rate}(A B) > \text{rate}(A NB)$	$\text{rate}(A B) < \text{rate}(A NB)$
<ul style="list-style-type: none">• Presence of A when B is present is stronger than B is absent• Positive association between A and B	<ul style="list-style-type: none">• Presence of A when B is present is weaker than when B is absent• Negative association between A and B

Basic rules on rates

- The overall $\text{rate}(A)$ will always lie between $\text{rate}(A|B)$ and $\text{rate}(A|NB)$
- The closer $\text{rate}(B)$ is to 100%, the closer $\text{rate}(A)$ is to $\text{rate}(A|B)$
- If $\text{rate}(B) = 50\%$, $\text{rate}(A) = \frac{\text{rate}(A|B) + \text{rate}(A|NB)}{2}$
- If $\text{rate}(A|B) = \text{rate}(A|NB)$, then $\text{rate}(A) = \text{rate}(A|B) = \text{rate}(A|NB)$

Symmetry Rule

$$\text{rate}(A|B) > \text{rate}(A|NB) \Leftrightarrow \text{rate}(B|A) > \text{rate}(B|NA).$$

$$\text{rate}(A|B) < \text{rate}(A|NB) \Leftrightarrow \text{rate}(B|A) < \text{rate}(B|NA).$$

$$\text{rate}(A|B) = \text{rate}(A|NB) \Leftrightarrow \text{rate}(B|A) = \text{rate}(B|NA).$$

To identify if there is any association, check for either :

1. $\text{rate}(A|B) \neq \text{rate}(A|NB)$ OR
2. $\text{rate}(B|A) \neq \text{rate}(B|NA)$

$\text{Rate}(\text{Success} | X) < \text{rate}(\text{Success} | Y) \rightarrow$ Negative association between successful treatment and X

Confounders

- A confounder is a third variable, associated with both dependent AND independent variable.
- If individuals are not randomly assigned to control and treatment, it is highly possible that the treatment and control groups have differences (confounder!)

Simpson's Paradox

- Relationship between rates in subgroups are reversed when subgroups are combined.
- SURE SIGN of confounding variable when encountering Simpson's Paradox.
- Confounders DOES NOT always lead to Simpson's Paradox.

If $R(A|B) > R(A|NB)$,

$R(B|A) > R(B|NA)$

$R(B|NA) < R(B|A)$ [Swap Side]

$R(NA|B) < R(NA|NB)$

$R(NA|NB) > R(NA|B)$ [Swap Side]

$R(NB|NA) > R(NB|A)$

$R(NB|A) < R(NB|NA)$ [Swap Side]

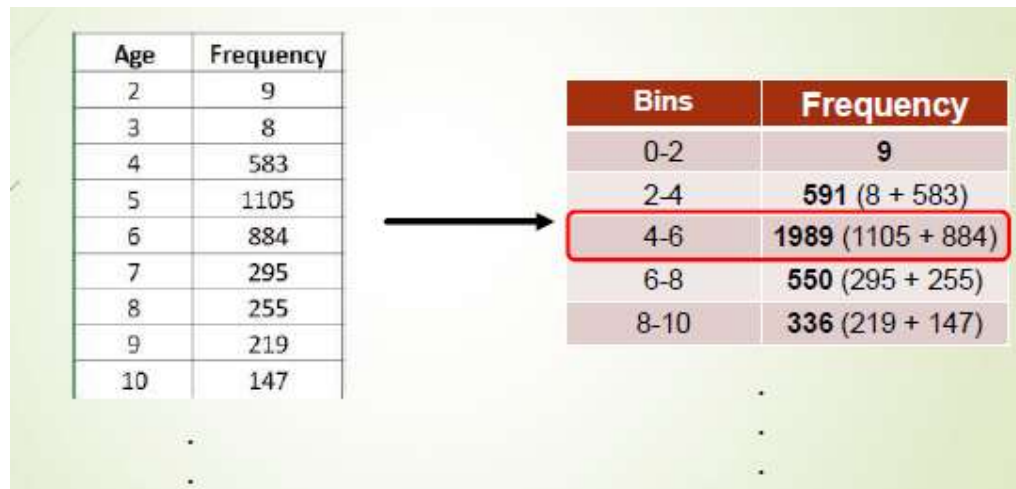
$R(A|NB) < R(A|B)$

$R(A|B) > R(A|NB)$ [Swap Side]

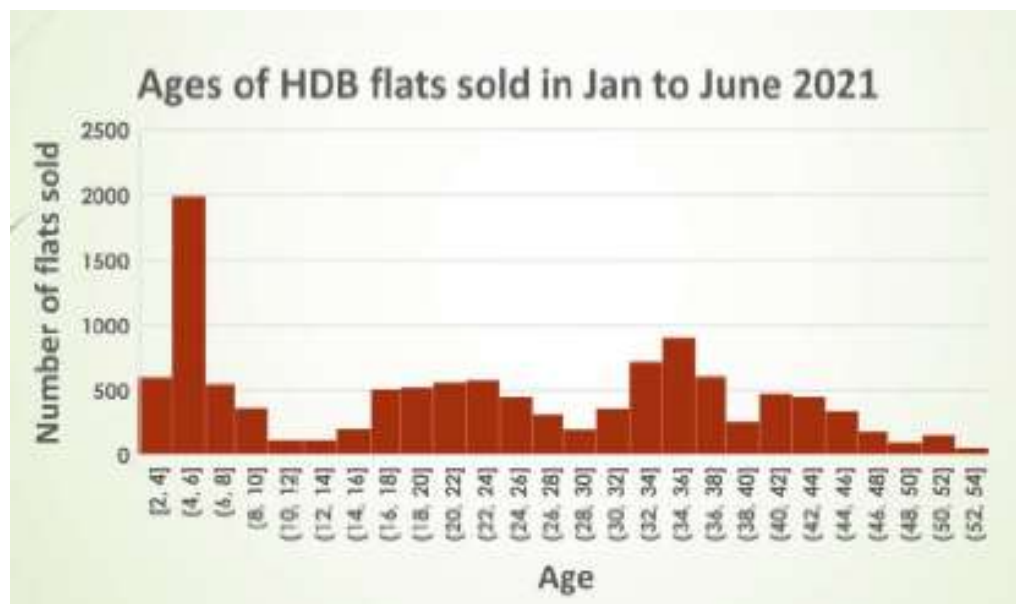
Chapter 3 – Dealing with Numerical data

3.1 Histograms

- Graphical display of a distribution
- Quick and easy to grasp
- Useful for large data sets



Values are divided into equal-sized intervals called bins.

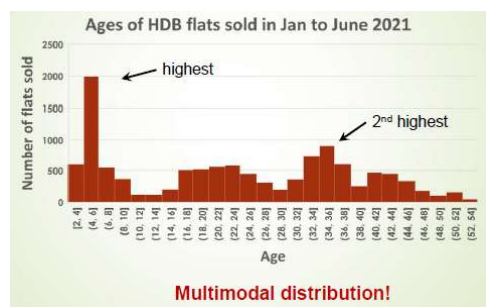


Describing distributions of Histogram

- Overall Pattern
 - o Shape
 - o Center
 - o Spread
- Deviations from the pattern
 - o Outliers

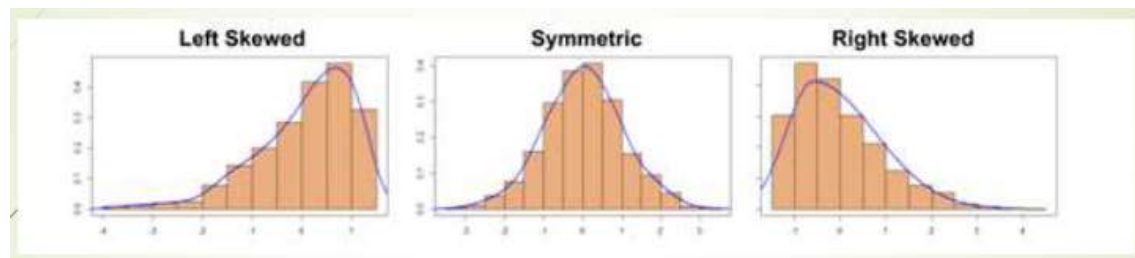
Shape of a distribution: Peaks and Skewness

Peaks



- Unimodal: One distinct peak
- Bimodal: Two peaks
- Multimodal: multiple peaks

Skewness

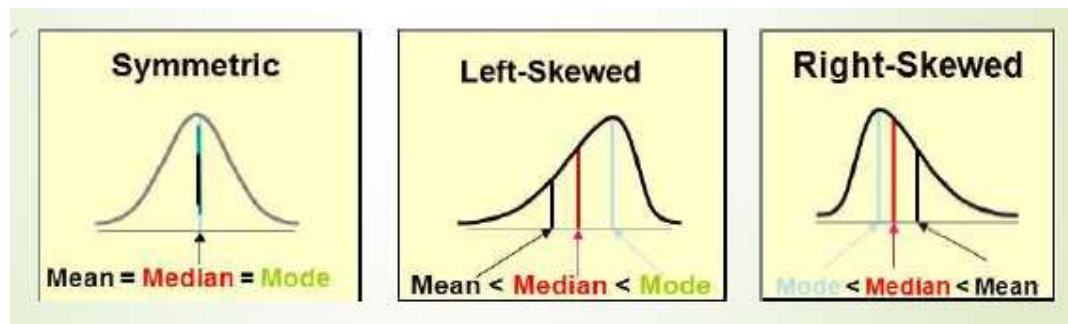


Another characteristic of the shape of a **(unimodal)** distribution is whether it is symmetrical or skewed.

Example of Symmetrical Distribution

- IQ Scores
- Bell Curve

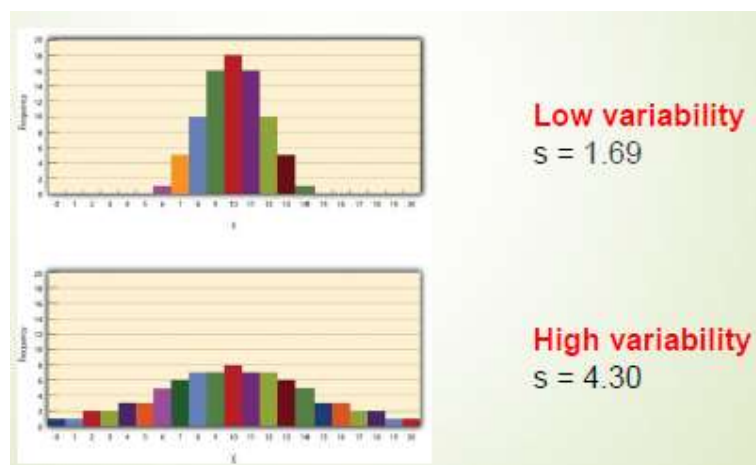
Center of a distribution: Mean, median, and mode



Three most common measures of central tendency: mean, median and mode.

- **Symmetric** : The **mean, median, and mode will be very close to each other** at the peak of the distribution (like the IQ example)
- **Left Skewed** : **Usually have mean < median < mode**
- **Right Skewed** : **Usually have mean > median > mode**

Spread of a distribution: Range & Standard Deviation



We can consider the variability or spread of a distribution when describing its pattern. That is how the data vary around their central tendency.

Most common measure of variability is the standard deviation.

Another simpler measure of variability is the range, which is the difference between the highest and lowest data points in the distribution.

- Range can be misleading -> Need to look at distribution

Outliers

- An outlier is an observation that falls well above or well below the overall bulk of the data.
- Examining data for outliers can be useful in
 - o Identifying strong skew in a distribution
 - o Identifying possible data collection or data entry errors
 - o Provide interesting insight into the data
- **It may be good practice to repeat an analysis of a data set with and without the outliers.** If they have minimal effect on the conclusion, and we can't figure out why they are there, we may possibly remove them. However, if they substantially affect the results, we should not drop them without justification.
- Data set should be sufficiently large to remove outliers

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

- We call the median and mode **robust statistics** if outliers have little or no effect on these values

Bin Size of Histograms

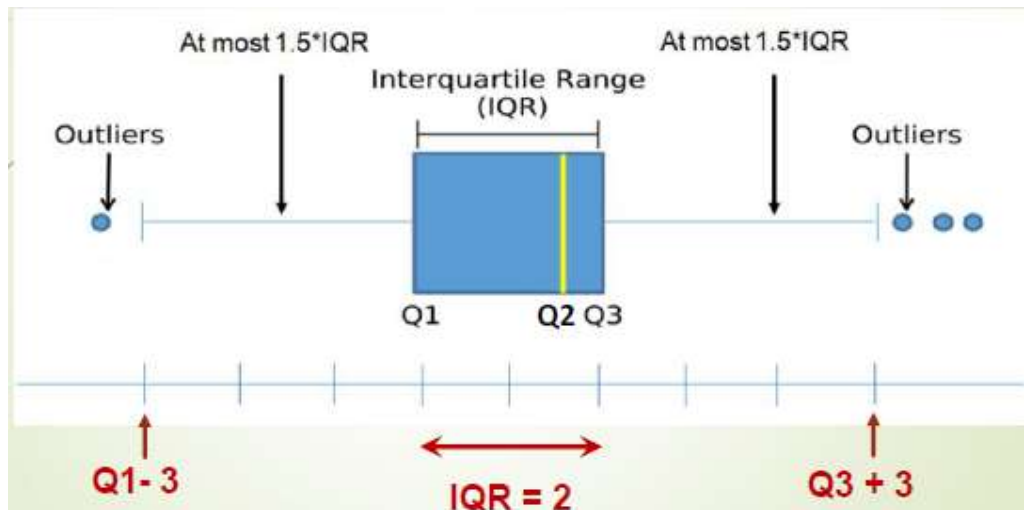
- **The bin size of a histogram matters**
- Avoid histograms with large bin widths that group data into only a few bins
 - o Does not give good information about variability in the distribution
- Avoid histograms with very small bin widths that group data into too many bins
 - o Doesn't give us a sense of the distribution
- **Construct histograms with different bin sizes to see which one is the most useful for our purpose**

Bar Graph vs Histograms

- **Histogram** shows the distribution of a **numerical variable** across a number line
- **Bar graph** makes comparisons across **categories (categorical variable)** of a variable
- The ordering of bars in a histogram cannot be changed, unlike bar graph
- Usually there are no gaps between bars in a histogram

Boxplots

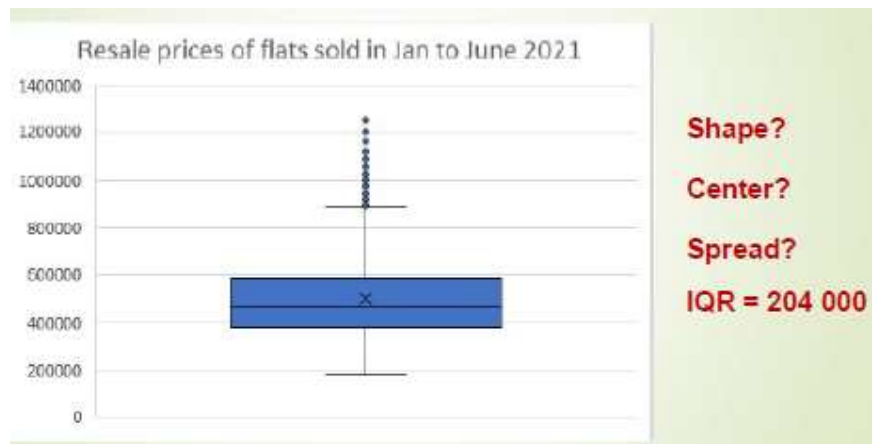
- Boxplot is another tool to visualise a distribution.
- Uses **five-number summary**:
 - o **Minimum**
 - o **Quartile 1 (Q1)**
 - o **Median (Q2)**
 - o **Quartile 3 (Q3)**
 - o **Maximum**
- **$IQR = Q3 - Q1$**
- Median can be thought of as a centre of the data set and the IQR is a way to quantify the spread of a data set
- **A data point is considered an outlier if it satisfies one of the following conditions**
 - o **Its value is greater than $Q3 + (1.5 * IQR)$**
 - o **Its value is less than $Q1 - (1.5 * IQR)$**



Steps to construct boxplot

- Draw a box from Q1 to Q3
- Draw a vertical line in the box where the median is
- Extend a line from Q1 to the smallest value that is not an outlier and from Q3 to the largest value that is not an outlier. These lines are called whiskers. Note that they may not be of the same length.
- Indicate outliers with dots.

Describing shape, centre, and spread using Histogram

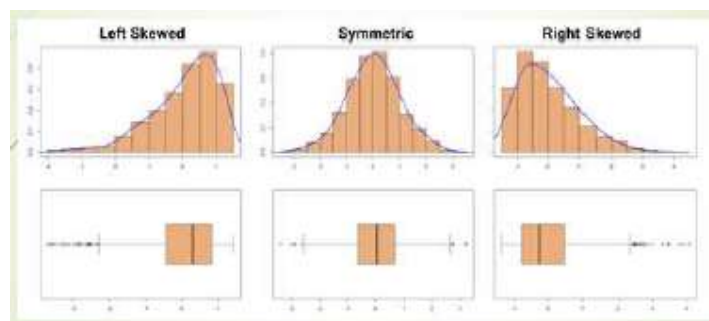


Shape: Compare the variability in the upper half of the data (Max – Median) to the variability in the lower half of the data (Median – Min). The distribution is skewed to the right because the lower half of the data has less variability than the upper half, and there is a relatively long tail to the upper end of the distribution due to the outliers.

Center: We can have a sense of where the median value is at a glance as compared to that of a histogram.

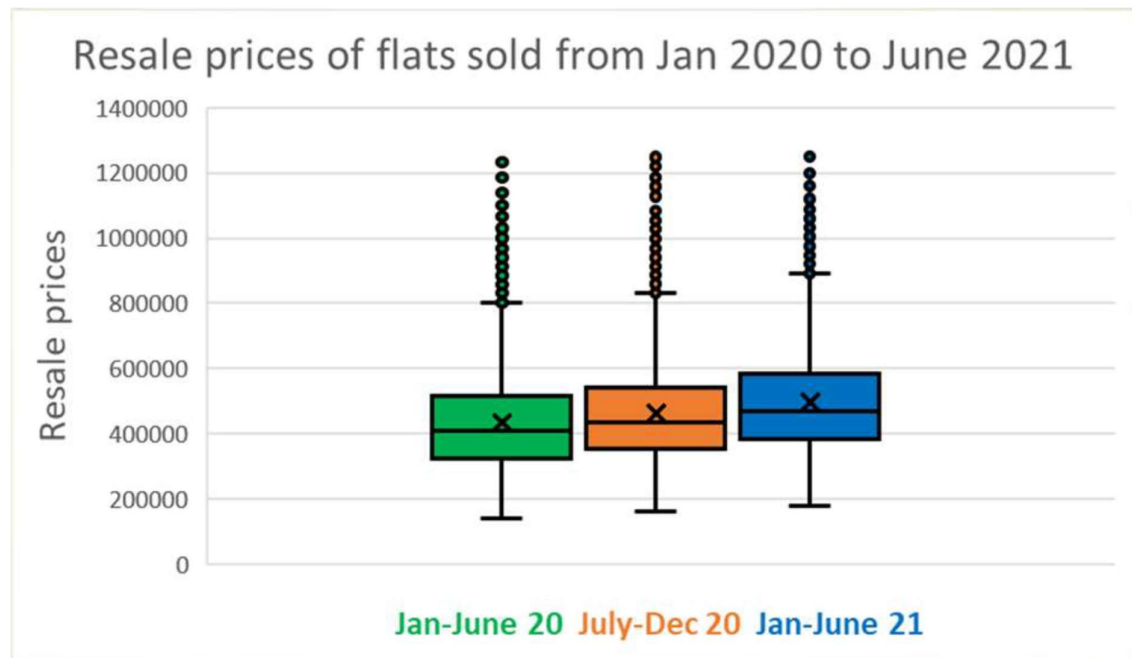
Spread: The IQR of 204 000 gives us an idea of the spread for the middle 50% of the data set. This would be a more meaningful measure if we were to compare the spread of different distributions in boxplots.

Boxplots vs Histograms



- Histograms can provide a better sense of the shape of distribution of a variable, especially when there are great differences among the frequencies of the data points.
- Boxplots are more useful when comparing distributions of different data sets.
- Boxplots can identify and exhibit outliers clearly.
- Boxplots do not give any information about how many data points we are working with. In other words, **two boxplots can look the same but correspond to data sets with very different numbers of data points.**

Comparing Boxplots



Shape – Shapes are similar (all right-skewed), but upper half of the data in earlier periods have greater variability.

Center – Median resale price have increased over time.

Spread – The IQR is 190000, 192000, and 204000 respectively.

Outliers – There are more outliers during the earlier periods. The largest outlier in all the periods is about the same but the lowest outlier in an earlier period is lower than the later ones. Perhaps this shows that the sale of expensive flats is not too affected by economic conditions?

Robust Statistics

IQR is a robust statistic. The IQR of a boxplot will remain the same even if outliers are removed.

Median is also a robust measure of central tendency.

Bivariate Exploratory Data Analysis

Deterministic Relationship

The value of one variable can be determined exactly if we know the value of the other variable.

Statistical Relationship

- Variability exists in measurements.
- Not possible to find a unique value of one variable corresponding to each value of the other variable
- Describe the average value of one variable, given the value of the other variable

Bivariate Data Analysis

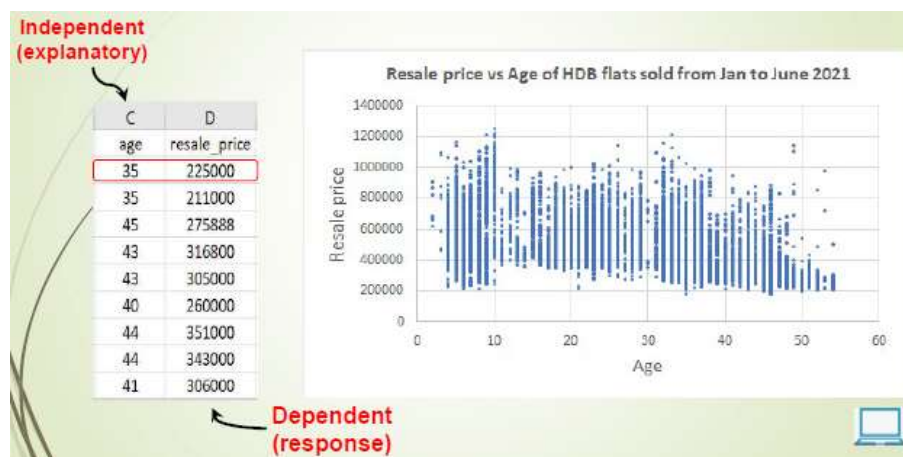
Scatter plots – to have an idea of the pattern formed between two variables.

Correlation coefficients – to check if the data are linearly related.

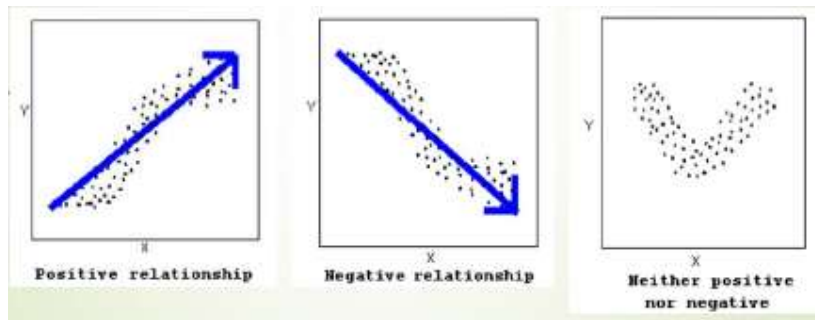
Regression analysis – to fit a line or curve to a data set and do predictions using the data.

Describing distribution of Scatterplot

- **Overall pattern**
 - o Direction
 - o Form
 - o Strength
- **Deviations from the pattern**
 - o Outliers



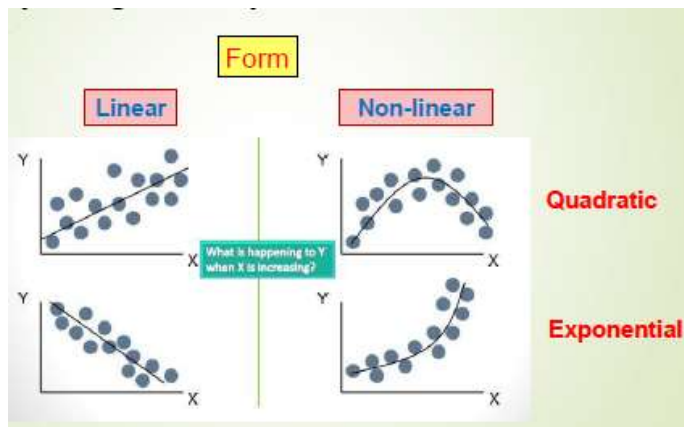
Direction of the relationship



The direction of the relationship can be positive, negative, or neither.

- A positive (or increasing) relationship means that an increase in one of the variables is associated with an increase in the other.
- A negative (or decreasing) relationship means that an increase in one of the variables is associated with a decrease in the other.
- Not all relationships can be classified as positive or negative

Form of the relationship

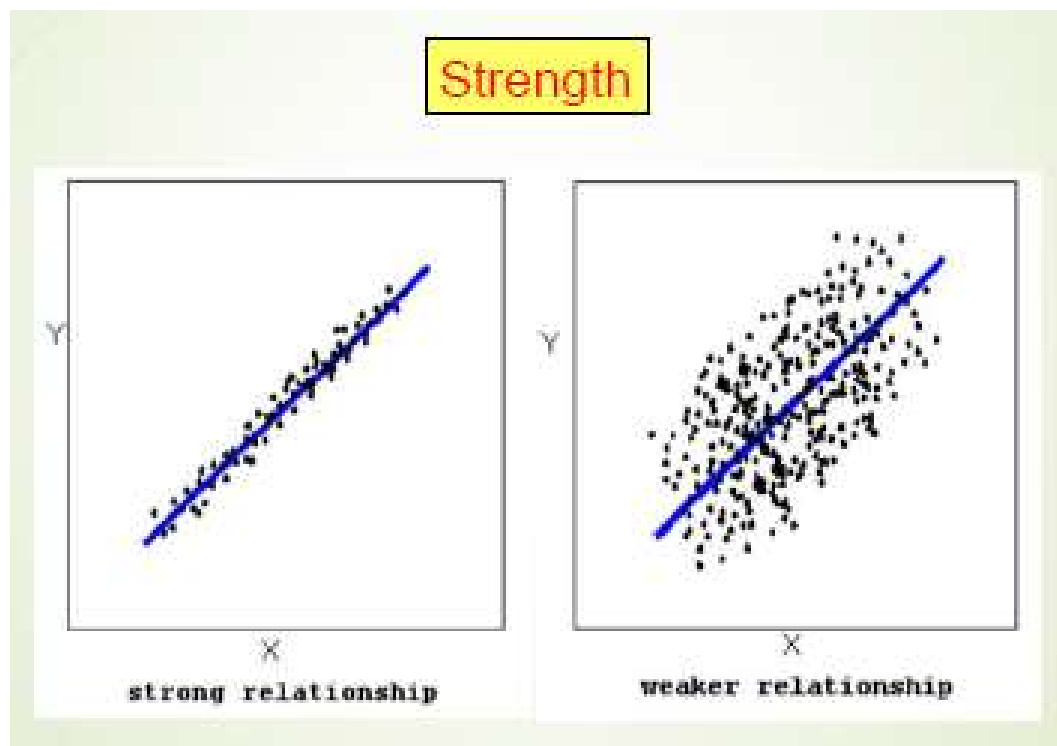


The form of the relationship is its general shape. Forms can generally be classified as linear or non-linear forms.

Linear form: The data points appear scattered about a line.

Non-Linear form: The data points appear scattered about a smooth curve. It is beyond the scope of this scope to summarise curved patterns.

Strength of the relationship



The strength of the relationship is a description of how closely the data follow the form of the relationship.

Outliers



Outliers are points that deviate from the pattern of the relationship.

Correlation Coefficient

- **Correlation coefficient is a measure of linear association**
- Range is between -1 and 1
- Summarizes the direction and strength of linear association

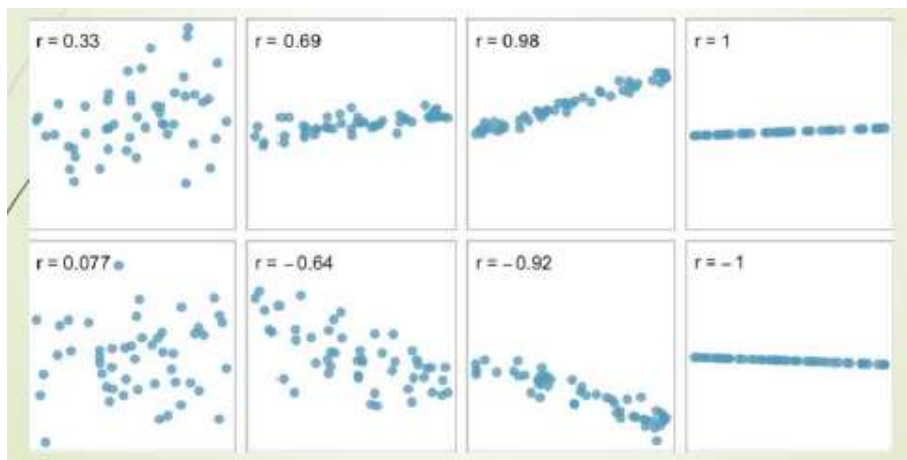
$r > 0 \rightarrow$ positive association

$r < 0 \rightarrow$ negative association

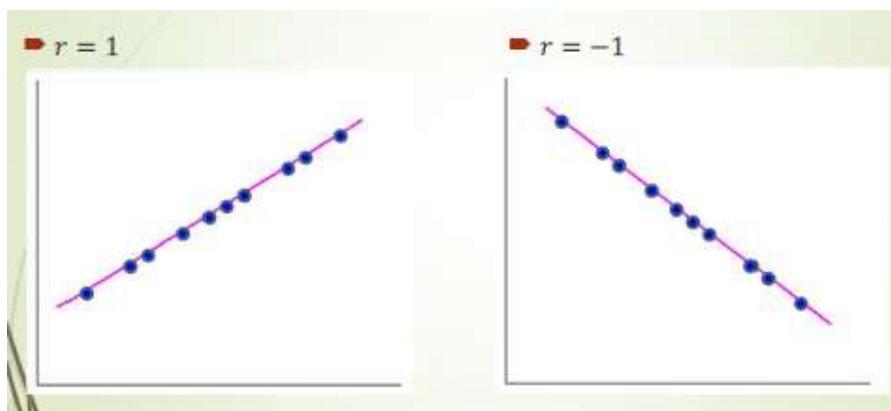
$r = 0 \rightarrow$ no linear association

$r = 1 \rightarrow$ perfect positive association

$r = -1 \rightarrow$ perfect negative association

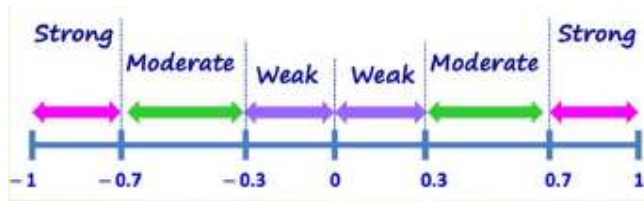


Examples



- $r = 1$, if the two variables are positively associated.
- $r = -1$, if the two variables are negatively associated
- $r = 0$, if it is a vertical or horizontal line. The change of value in one of the variable did not cause a change in the other variable. Thus there is no association.

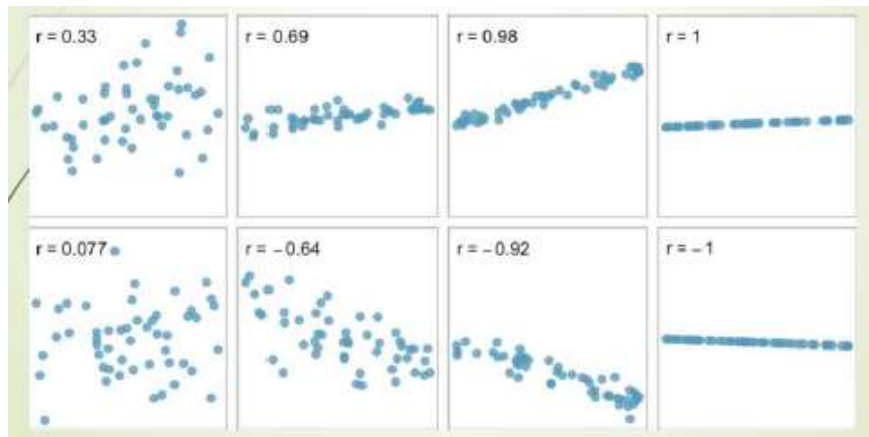
Interpreting r value



(Rule of thumb)

- Magnitude of r tells us about the strength of linear association
- The closer the value of r is to 1 or -1, the stronger the linear association.
- The closer the value is to 0, the weaker the linear association.

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



How to compute correlation coefficient?

- Convert each data point into its standard unit
- $SU_X = \frac{X - \text{average}(X)}{s_x}$, $SU_Y = \frac{Y - \text{average}(Y)}{s_y}$, where s_x is S.D. of X
- Sum the product of each data point (in S.U.), and find the average
- r value is just the average of the product of X and Y in standard units
- Note: Not expected to compute r value by hand

Example

$$X = 9 \rightarrow \frac{9 - 5.5}{2.87} = 1.22$$

$$Y = 41 \rightarrow \frac{41 - 25.1}{14.84} = 1.07$$

X	Y		
9	41	Average of X	5.5
4	17	Average of Y	25.1
5	28	Standard deviation of X	2.87
10	50	Standard deviation of Y	14.84
6	39		
3	26		
7	30		
2	6		
8	4		
1	10		

X	Y		X (standard unit)	Y (standard unit)	Product
	9	41	1.22	1.07	1.31
	4	17	-0.52	-0.55	0.29
	5	28	-0.17	0.20	-0.03
	10	50	1.57	1.68	2.63
	6	39	0.17	0.94	0.16
	3	26	-0.87	0.06	-0.05
	7	30	0.52	0.33	0.17
	2	6	-1.22	-1.29	1.57
	8	4	0.87	-1.42	-1.24
	1	10	-1.57	-1.02	1.59

$$r = \frac{1}{10} (1.31 + 0.29 + \dots + 1.59) = 0.64$$

Properties of r

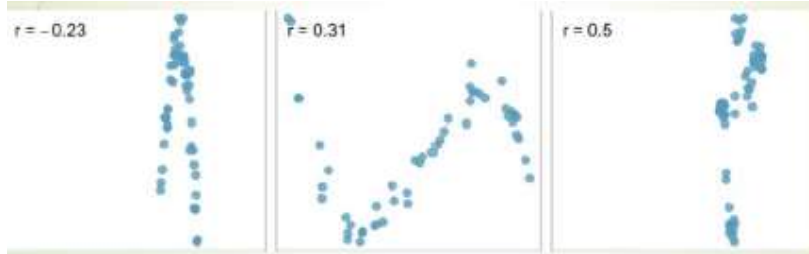
r is not affected by the following operations.

- Interchange of two variables (Interchanging x and y axis)
- Adding/Subtracting a number to all values of a variable
- Multiplying/Dividing a **positive** number to all values of a variable

Limitations of Correlation Coefficient

- **Correlation does not imply causation!!**
 - o We can only conclude there is a **statistical relationship**, but **not a casual relationship**
- There might be a third variable correlated to the two variables (confounders)

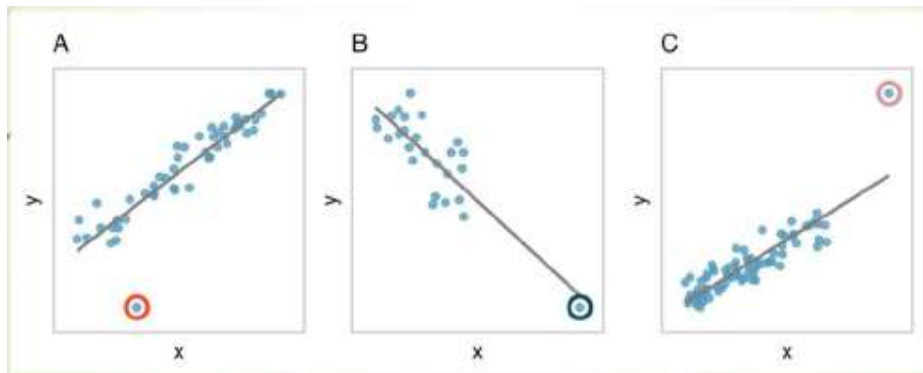
Non-linear association



- r only measures linear association between two variables
- Always look at the scatter plot and not only at the r value

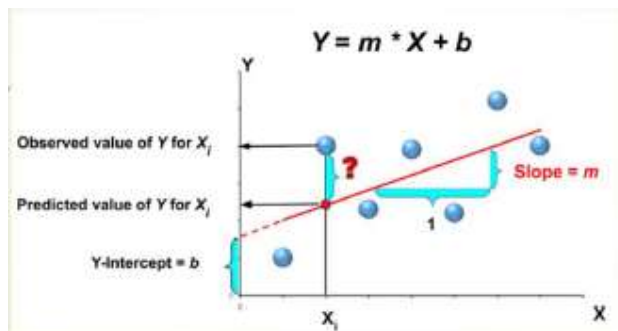
There may be a strong relationship between two variables, but the r value is small because the association between the two variables is not linear.

Outliers



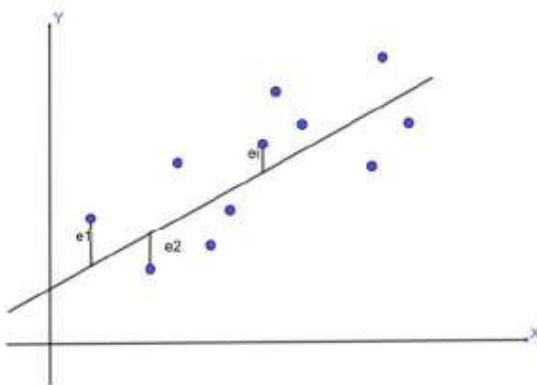
- Outliers can affect the strength of linear association between two variables.

Linear Regression

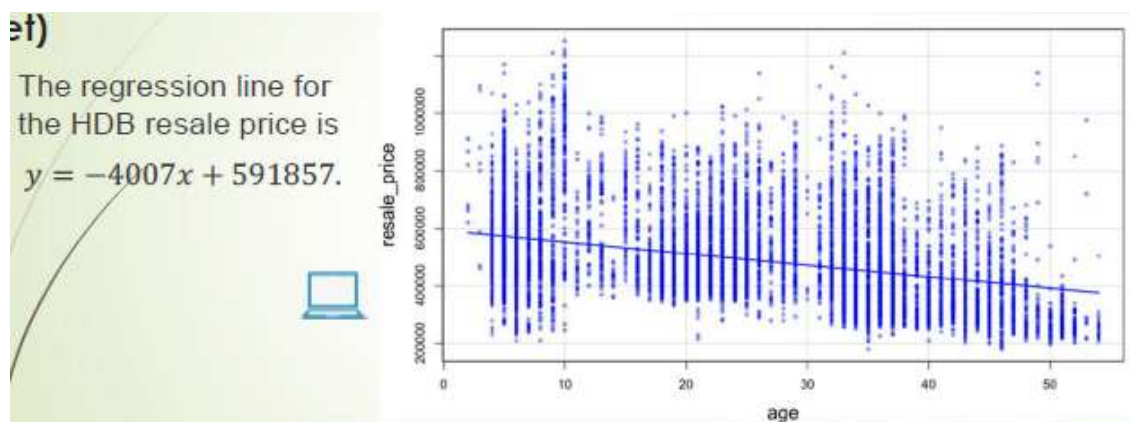


- We model the relationship by a straight line
- $Y = mX + b$
- Constant b is the y-intercept
- Constant m refers to the gradient of the line

How to find regression line



- Define the i -th residual of the observation: e_i = difference between the observed outcome and predicted outcome.
- Want to minimize $e_1^2 + \dots + e_n^2$
- Least square method is used to determine regression line



- The equation can only be used for predicting HDB resale price given its age. It CANNOT be used to predict the age of an HDB flat given its resale price.

Slope vs Correlation Coefficient

The slope of the regression line and correlation coefficient is related by the equation,

$$m = \frac{s_y}{s_x} r \quad \text{or} \quad r = \frac{s_x}{s_y} m$$

where s_y is the S.D. for y , and s_x is the S.D. for x .

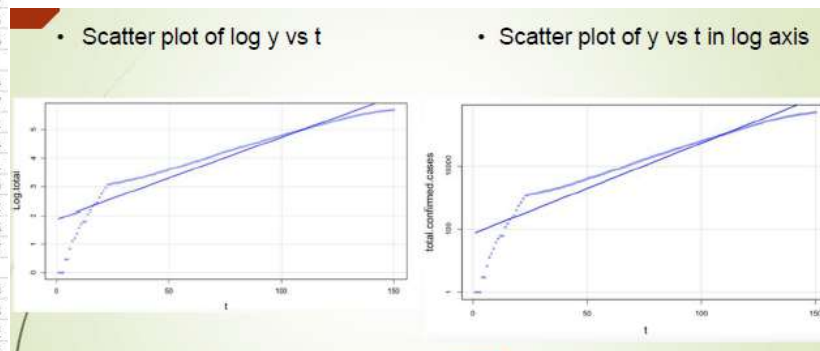
- The slope of the regression line is not necessarily equal to the correlation coefficient.
- If the correlation coefficient is positive, then the gradient is also positive. Likewise for negative.

Extrapolation

Extrapolating beyond the observed range is dangerous, as the best fit regression line may change.

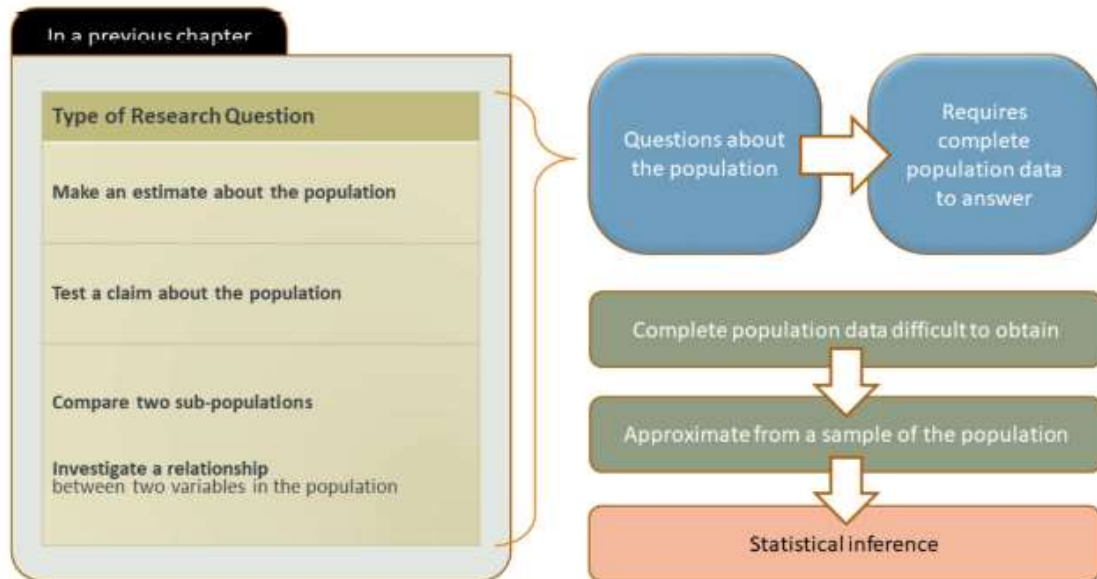
Modelling exponential curve

t	total confirmed cases	Log total
76	17200	4.2355284
77	18003	4.2553449
78	19137	4.2818739
79	20125	4.3037359
80	21343	4.3292555
81	22583	4.3537816
82	23615	4.373188
83	24264	4.3849624
84	25937	4.4139197
85	27403	4.4377981
86	29240	4.4659774
87	30967	4.4908991
88	32683	4.5143219
89	34357	4.5360152
90	35812	4.5540286
91	37525	4.5743207
92	40792	4.610575
93	43434	4.6378298
94	45973	4.6625028
95	48285	4.6838122
96	50879	4.7065386

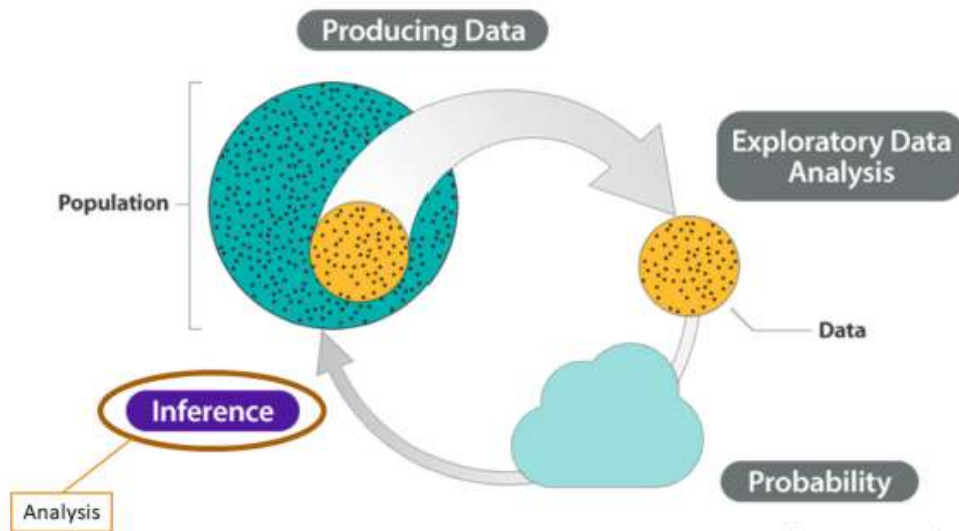


- Model the relationship above as $y = cb^t$
- Find c and b using linear regression model
- **First Step: plot log y vs t**
- **Second Step: find regression line for log y vs t**
- We find $\log y = 0.0286t + 1.86$
- **Last Step: Express y in terms of t**
- $y = (10^{1.86}) (10^{0.0286t})$

Chapter 4 - Statistical Inference



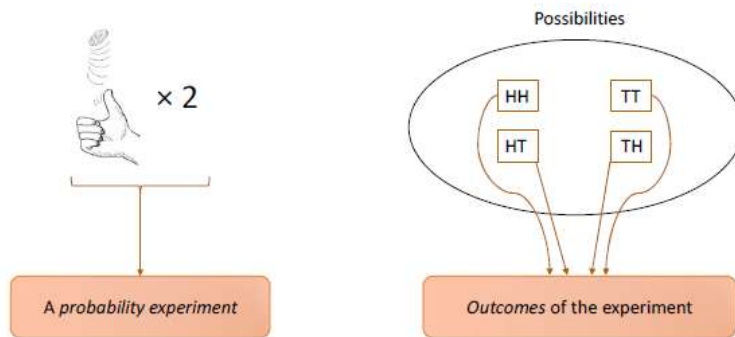
The common thread among all the types, is that they are all **questions pertaining to the population**. **Answers to such questions ideally requires complete knowledge about the population**, at least with respect to the relevant characteristic. This is often untenable, because of **time and logistical constraints of the real-world**. It is therefore, useful to be able to give an **approximate answer based on just a sample of the population**. The general approach of drawing conclusions about the population from sample data is known as statistical inference.



Picture source: courses.lumenlearning.com

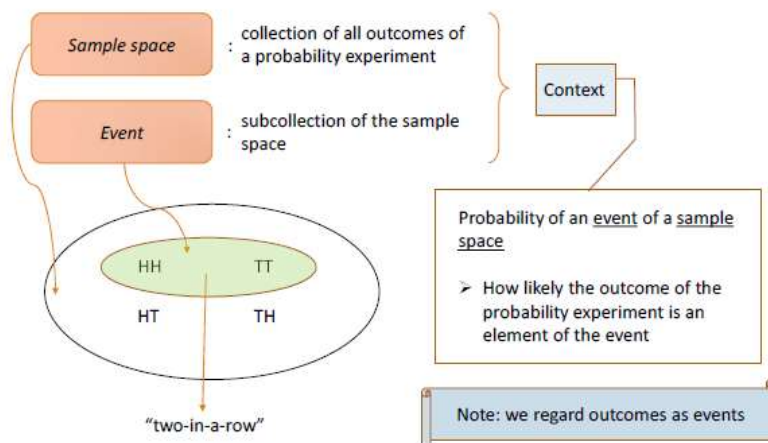
The end goal of the “analysis” step in PPDAC is very often statistical inference.

Probability



A probability experiment must:

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

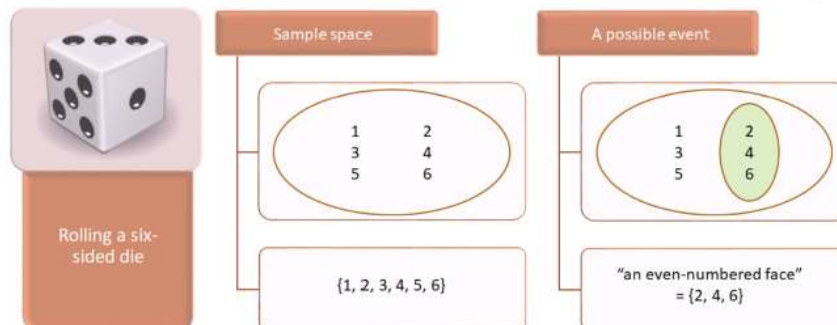


Sample space: collection of all outcomes of a probability experiment

Event: subcollection of the sample space. (Eg. 2H or 2T -> two-in-a-row)

Probability of an event of a sample space: How likely the outcome of the probability experiment is an element of the event

Example: Die-rolling



Probabilities: Numerical values between 0 and 1 (inclusive)

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

Simple Cases: Finite Sample Spaces

For any event E :

1. Repeat the probability experiment a large number (N) of times
2. For each repetition, check if the outcome is in E

Every event can be assigned a probability

- Proportion of E estimates the true $P(E)$
- Estimate gets more accurate as N increases

Proportion of $E = \frac{\text{Count of Yes}}{N} \rightarrow P(E)$

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 sample spaces, it is enough to assign probability to outcomes so that they add up to 1.

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

$P(1) = 0.1$
 $P(2) = 0.1$
 $P(3) = 0.1$
 $P(4) = 0.1$
 $P(5) = 0.1$
 $P(6) = 0.5$

add up to 1

Deriving probabilities of other events

E.g. Let E denote the event “an odd-numbered face”.
Let F denote the event “an even-numbered face”.

$P(E)$ $= P(1) + P(3) + P(5)$ $= 0.1 + 0.1 + 0.1$ $= 0.3$	$P(F)$ $= P(2) + P(4) + P(6)$ $= 0.1 + 0.1 + 0.5$ $= 0.7$
--	--

Rolling a biased six-sided die

Uniform Probabilities and Rates

In the case of finite sample spaces, uniform probability assigns equal probability to every outcome. Each outcome has the probability one divided by size of sample space.

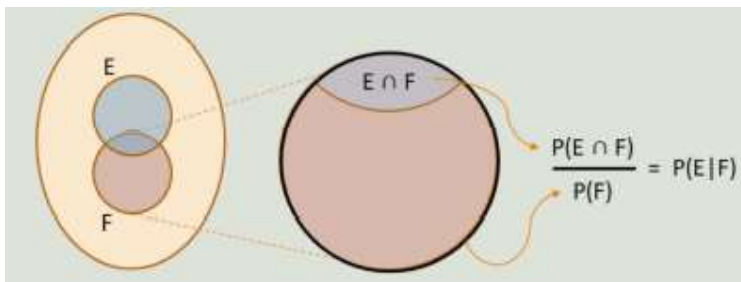
Uniform probability \rightarrow Every outcome has same probability $= \frac{1}{\text{size of sample space}}$

Uniform Probability & 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.
- **For any subgroup (event) A, $P(A)$ = probability of select unit being in A = rate(A)**

Conditional Probability

$P(E | F)$: “probability of E given F”; How likely the outcome is in E, if we know it is in F.



To compute $P(E | F)$: we restrict our focus to the given event F, which may contain some overlap with E, denoted “E intersect F”.

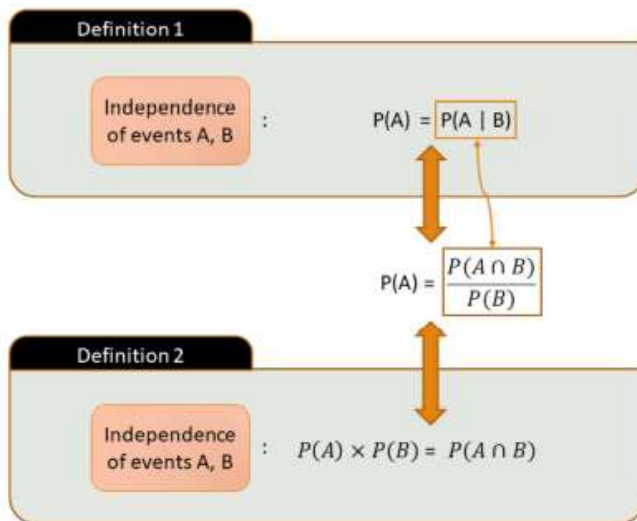
- If $P(F) == 0$, by convention, $P(E | F) = 0$ for all events E of the same sample space as F.

? Does $P(A | B)$ equal rate(A | B)?

Yes!

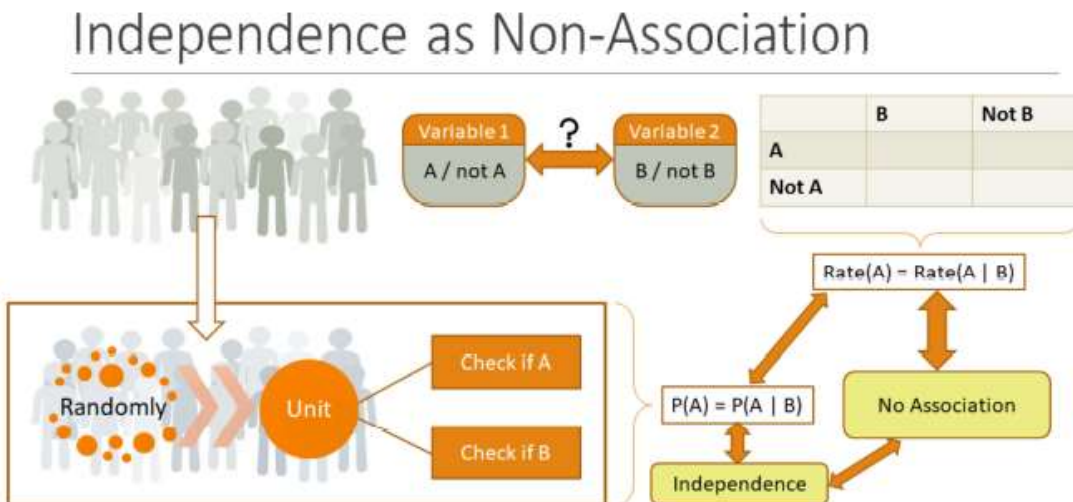
$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$
$$= \frac{\text{rate}(A \cap B)}{\text{rate}(B)}$$
$$= \frac{\frac{\text{size of } A \cap B}{\text{size of sampling frame}}}{\frac{\text{size of } B}{\text{size of sampling frame}}} = \frac{\text{size of } A \cap B}{\text{size of } B} = \text{rate}(A | B)$$

Independence

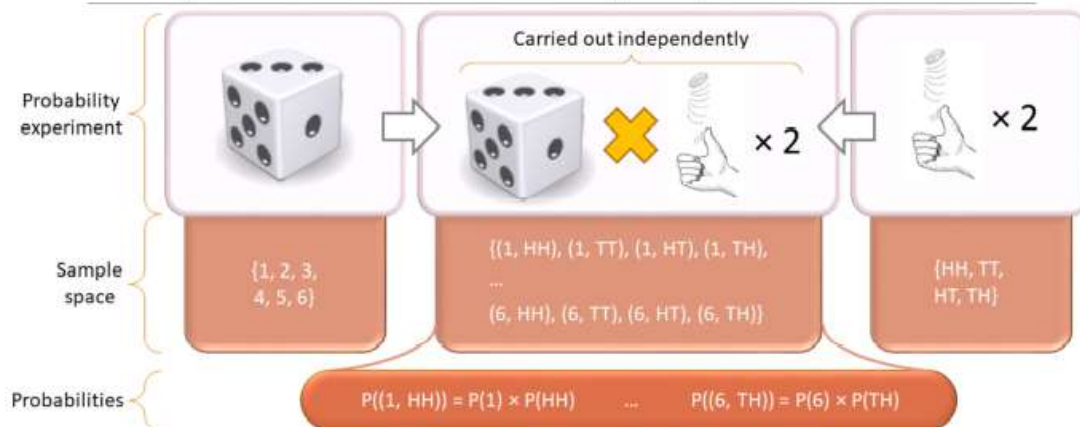


- Order of events does not matter when we talk about independence

Independence as non-Association



Independent Probability Experiments

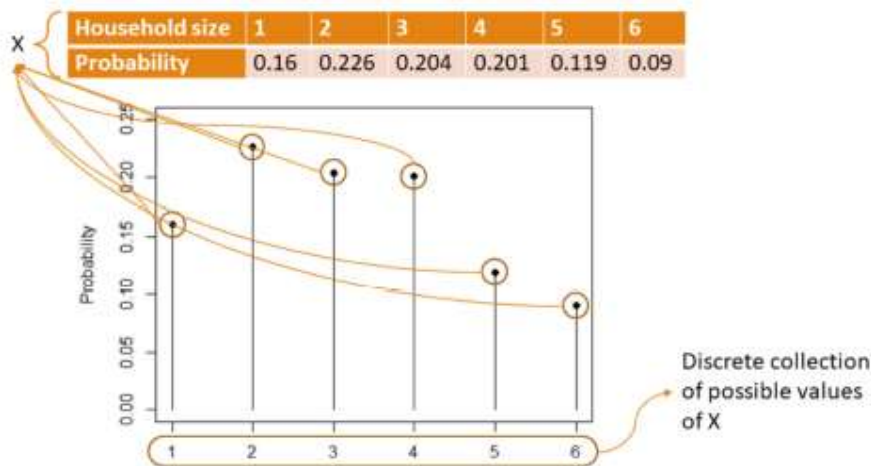


- Pairings obey "Definition 2" of independent events

Random Variables

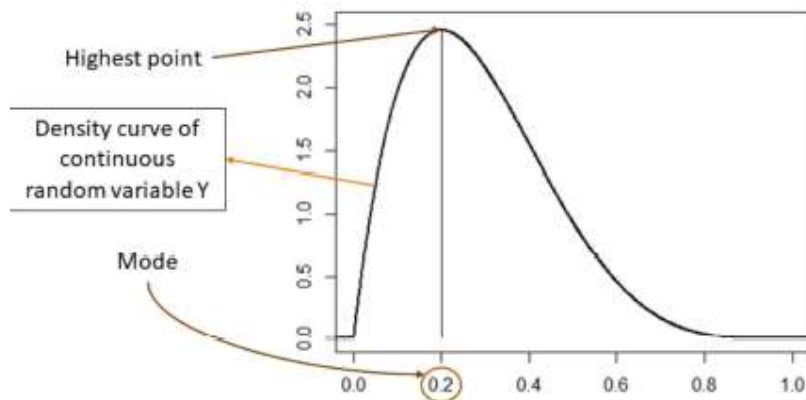
- Any numerical variable with probabilities assigned over its possible values, is a random variable
- If the numerical variable is a discrete variable, we call the random variable a discrete random variable. If the numerical variable is a continuous variable, we call the random variable a continuous random variable.

Visualization of Discrete Random Variable



- Probabilities of the points add up to 1
- X-value of a highest point is a mode
- $P(X \geq 5) = P(5) + P(6) = 0.119 + 0.09 = 0.209$

Visualization of a Continuous Random Variable



- Area under the curve = 1
- X-value of a highest point is a mode
- $P(0.3 \leq Y \leq 0.5) = \text{Area under curve from 0.3 to 0.5}$
- **Probability that a continuous random variable takes on a value in interval $[a,b]$ = area under its density curve from a to b**

Normal Distribution

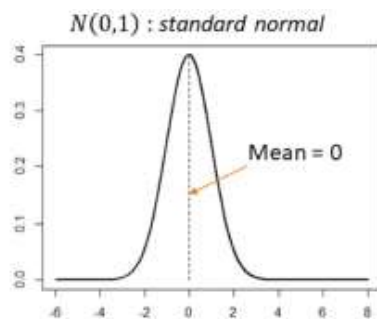
Two normal distributions can only differ by their means or their standard deviations

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

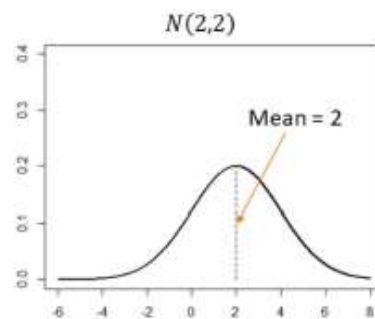
SD = \sqrt{y}

Common Properties:

- Bell-shaped curve
- Peak of the curve occurs at the mean (Mean = mode = median)
- Curve is symmetrical about the mean (Mean = mode = median)

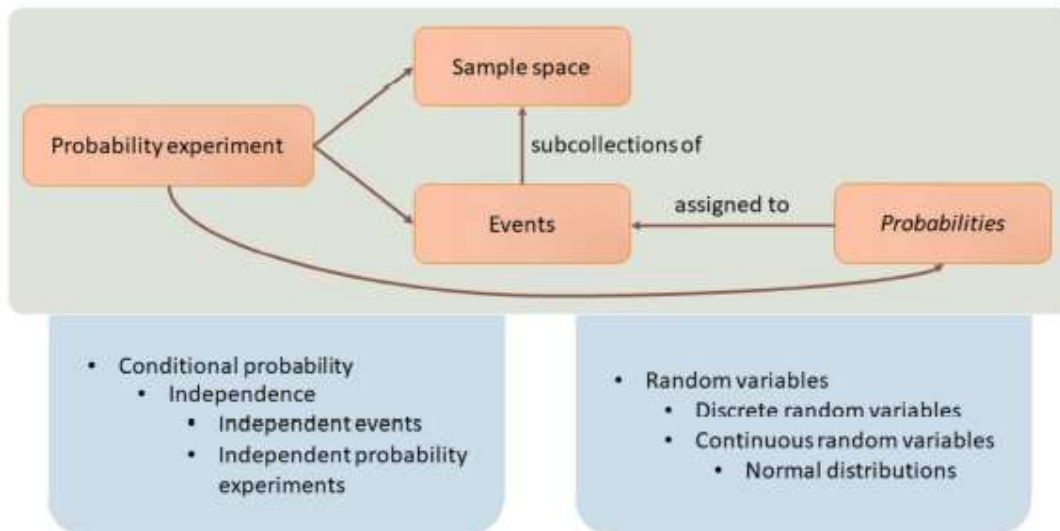


• Smaller standard deviation → thinner bell shape



• Greater standard deviation → fatter bell shape

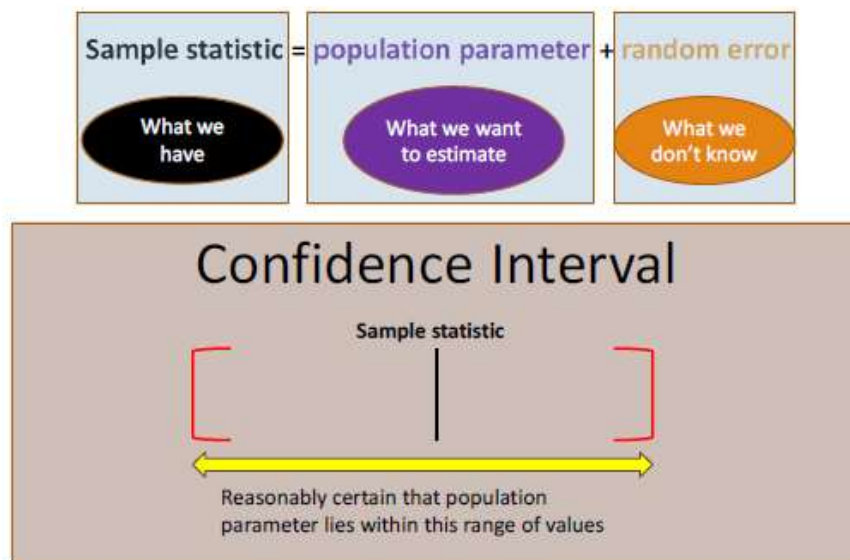
Area under curve kept constant → a fatter curve compensates by being shorter



Statistical Inference

- In a probability sample, we only need to deal with random error. We will assume that the samples are simple random sample from a perfect sampling frame, with 100% response rate.
- Sample statistic = population parameter + random error
- Since the samples are random, different samples have different sample means, so we can use probability to say how confident we are about our conclusions regarding the population from which the samples are drawn.

Confidence Intervals



0.16 = population proportion of 1-member households + random error

Sample proportion $p^* = 0.16$

Random sample of 1000 households from Singapore

Number of individuals living in a household	1-member household	2-member household	3-member household	4-member household	5-member household	6-member household
Number of households	160	226	204	200	119	91
Proportion	0.16	0.226	0.204	0.2	0.119	0.091

Value from standard normal distribution z^*

- For 90% confidence interval, z^* is 1.645
- For 95% confidence interval, z^* is 1.96

Confidence interval for population proportion

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

Sample proportion, p^*

Sample size, n

Value from standard normal distribution, z^*

Confidence interval for population proportion of 1-member household:

95% CI: 0.16 ± 0.02

year	household	household	flat_price
2020	3-member Executive		800011.7
2020	3-member 5-room		684905.9
2020	4-member 4-room		578144.7
2020	5-member 4-room		648816.2
2020	4-member 4-room		573459.3
2020	4-member 5-room		717996.6
2020	4-member 5-room		612177.1
2020	1-member 3-room		488277.2
2020	3-member 4-room		565517.4
2020	2-member 3Gen		796074.1
1980	6-member 2-room		239556.7
2000	6-member 3Gen		738475.8
2020	3-member 3-room		509605.9
2020	4-member 2-room		337131.5
2020	1-member 5-room		707953.4
2020	4-member 3Gen		758605.9
2020	1-member 2-room		488993.7
2020	1-member 4-room		552096
2020	4-member 3Gen		784334.9
1980	6-member 4-room		458955.7
1980	1-member Executive		830863.8
2020	5-member 2-room		380128.4
2020	5-member 3Gen		731226.6
2020	4-member 3-room		533506.8
2020	5-member 4-room		590034.7
1980	6-member Executive		813863
2020	5-member 5-room		622018.7
1980	6-member 3Gen		699042.3
2020	5-member 3Gen		722737.6
1980	3-member Executive		777924.1
2020	1-member Executive		808943.7

Year	1980	2000	2020
Number of households	855	1660	2485



Sample mean flat price for year 2020, $\bar{x} = \$616,013$

Population mean flat price for year 2020, μ

616013 = population mean flat price for year 2020 + random error

Confidence interval for population mean

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

Sample mean, \bar{x}

Sample size, n

Sample standard deviation, s

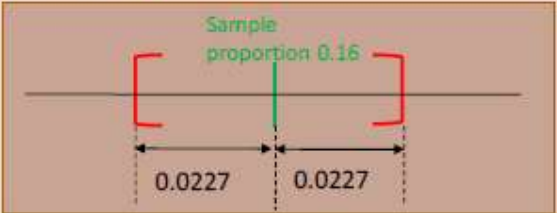
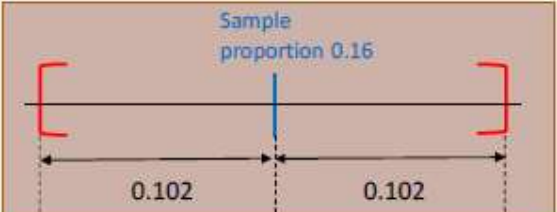
Value from t-distribution, t^*

To construct confidence intervals for population means, we require

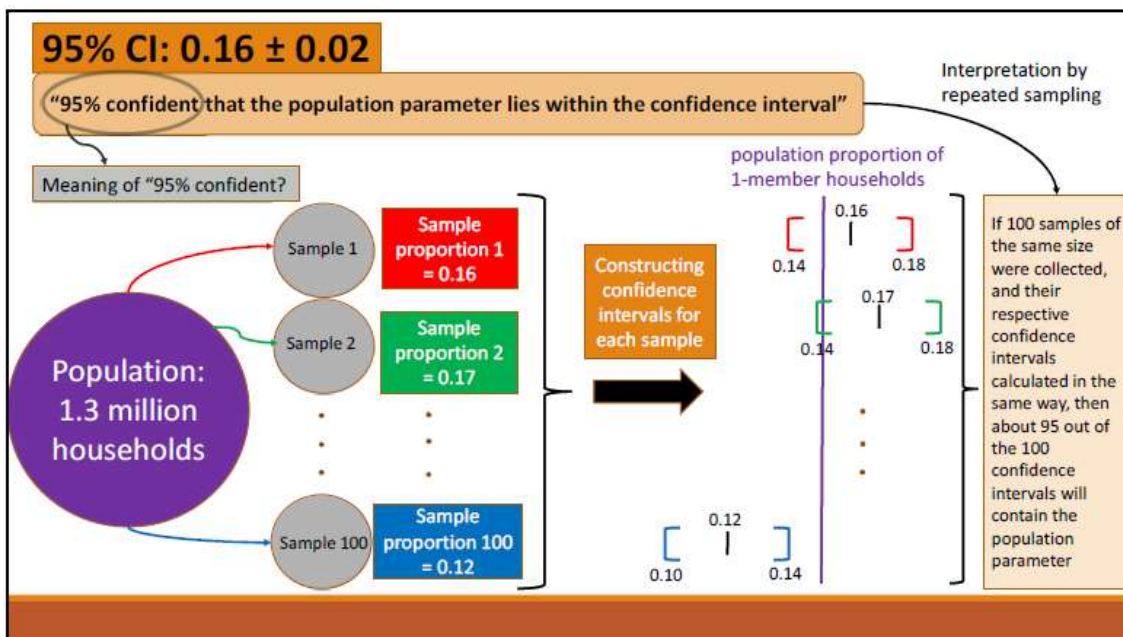
- Standard deviation of sample
- Size of sample
- Confidence level

Properties of confidence intervals

- The smaller the sample size, the larger the random error
- The higher the confidence level, the larger the random error

Sample size	Sample proportion	Confidence interval
1000	0.16	95% CI: 0.16 ± 0.0227 
50	0.16	95% CI: 0.16 ± 0.102 

Smaller sample imply there is a larger random error, which implies that the CI of the smaller sample will be larger when compared to the CI of the larger sample.



Hypothesis Testing

Sometimes we are face with a (yes/no) decision problem involving a large population. For example, is a vaccine effect and safe enough to be rolled out to the entire populace?

There are a few key steps to hypothesis testing

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

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

Step 3: Determining the level of significance (usually 5% level of significance) and computing the p-value.

Step 4: Making conclusion about the null hypothesis

Step 1: Formulating hypothesis

- Produce 2 hypotheses – **null hypothesis** and **alternative hypothesis**
- Investigates a proposed mathematical model to determine whether there is sufficient evidence from a sample of data to reject the defined null hypothesis
- The **null hypothesis** takes a stance of no difference or no effect. This hypothesis assumes that any differences seen are due to variability inherent in the population and could have occurred by random chance.
- The **alternative hypothesis** is typically what we wish to confirm and pit against the null hypothesis. We wish to reject the null hypothesis in favor of the alternative.
- The two hypotheses must be **mutually exclusive**; they cannot both be true.

Example

We suspect a coin is biased towards heads.

Null Hypothesis H0: “The coin is fair.”

Alternative Hypothesis H1: “The coin is biased towards heads.”

Step 2: Collect relevant data

- Collect relevant sample data for the experiment
- For the example, we can toss the coin 8 times and observe all the outcomes.

Our **random variable** is the number of heads out of 8 independent coin tosses, assuming coin is fair.

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

Step 3: Significance level and p-value

Significance level

- The lower the significance level, the greater the evidence needs to be 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)

Intuitively, observing 5 heads out of 5 coin tosses is a greater evidence of the alternative being more plausible than the null. We expect the former observation to pass a test of a fixed significance level than the latter more easily.

Computing the p-value

- **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.

Example: Observed 7 heads out of 8 tosses

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

Step 4: Making conclusion

Decide between one of the ONLY two options:

- Reject the null hypothesis in favor of the alternative if $\text{p-value} < \text{significance level}$
- Do not reject the null hypothesis, if $\text{p-value} \geq \text{significance level}$. Our test result is **inconclusive**.

The above example had a p-value of 0.035, we can reject the null hypothesis at the 5% level of significance ($0.035 < 0.05$) but not at the 3% level of significance.

Hypothesis Testing: Warning

- 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 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.

Common hypothesis tests

- How do we go about deciding on which test statistic to use?
- Is there any type of hypothesis test that we would like to use?

These decisions are made based on:

- The kinds of hypotheses we want to test
- The distributions of the test statistics

Conducting t-test

Criteria for conducting a t-test:

- The population distribution should be approximately normal if n , the sample size, is smaller than 30.
- The data used is produced randomly.

Chi-squared test

- Commonly used to check whether two categorical variables, A and B are significantly associated.

Criteria for conducting a chi-squared test:

- The data must be counts for the categories of a categorical variable.
- We would like the cases within each group to be selected randomly

Common Hypothesis Tests : Summary

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.