GEA1000 Notes

November 22, 2024

1 Chapter 1

1.1 Quantitative Reasoning with Data

1.1.1 Population

The population is the entire group that we want to know something about

1.1.2 Research Question

The research question seeks to investigate some characteristic about a population

1.2 Sampling

1.2.1 Samples

A proportion of the population selected in the study. Not feasible to gather information from every member of the population
Estimates are inferences based on information from the sample.

1.2.2 Sampling Frames & Generalisability

A sampling frame is the list from which the sample is obtained. The sampling frame must be well-picked in order to ensure *generalisability* to the population as whole.
The sampling frame must minimize redundant data and maximize coverage at the same time.
For higher generalizability:

1. Sampling frame equal to or larger than the population
2. Adopt probability based sampling to minimize selection Bias
3. Have a large sample size to reduce variability or random error in the sample
4. Minimize non-response rate

1.2.3 Bias

Avoid **selection bias** and **non-response bias**.
Selection Bias is the biased selection of units. Randomized sampling avoids this.
Non-response bias occurs when to-be participants avoid due to inconvenience, lack of interest, privacy, etc.

1.2.4 Census

A census is an attempt to reach out to the **whole** population.

1.2.5 Probability Sampling

Sampling process via a known randomised mechanism.
Probability of selection may not be the same throughout all units of sampling frame
Element of chance eliminates biases associated with selection.

1.2.6 Simple Random Sampling

Units selected randomly **without replacement** from sampling frame using random number generator. Every unit has equal chance to be selected.

1.2.7 Systematic Sampling

- Applying a selection interval K, and random starting point from first interval.
- Simple Process
- Not good with representation

1.2.8 Stratified Random Sampling

- Population broken down into strata of different size
- Units within strata share similar characteristics
- Characteristics vary across different strata
- Simple Random Sampling is done within every stratum.
- Good representation
- Need information about sampling frame and strata

1.2.9 Cluster Sampling

- Units broken down into similar clusters
- Randomly sample fixed number of clusters
- All observations within selected clusters are recorded
- Less tedious & time consuming
- Costly
- High variability if clusters are dissimilar

1.2.10 Non-Probability Sampling

- The non-usage of chance in selection of individuals
- Sampling methods are not mutually exclusive

1.2.11 Convenience Sampling

- Selection of subjects based on proximity and availability
- e.g. Mall Surveys
- Open to **selection bias** and **non-response bias**

1.2.12 Volunteer Sampling

- Researchers seek volunteers to participage
- e.g. Online Polls
- Open to **selection bias** and **non-response bias**

1.3 Variables and Summary Statistics

1.3.1 Independent and dependent variables

An independent variable is a variable that may be subject to manipulation in a study

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

1.3.2 Categorical Variables

Categorical values take label values, and each obs can be placed in only one label.

Ordinal Categorical
Ordinal Categorical variables come with some natural **ordering** and numbers are often used to represent.

Nominal Categorical
There is no intrinsic ordering for the variables - they're simply labels.

1.3.3 Numerical Variabels

Take numerical values. Arithmetic operations on these variables make sense.

Discrete Numerical
Discrete Numerical values are one where possibles values of the variable form a set of numbers with gaps. e.g. number of people

Continuous Numerical
Continuous Numerical values can meaningfuly take on all possible numerical values in a given range or interval.

1.3.4 Summary Statistics

Mean

- The mean is given $\bar{x} = \frac{x_1+x_2+...+x_n}{n}$
- Adding a constant value to all datapoints will change the mean by that constant
- Multiplying c to all datapoints will result in mean being multiplied by c

Variance

- Given $\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n-1}$
- Tells us absolute spread of data in squared units
- Adding a constant value to all datapoints will not change the variance value.
- Multipliying data points by c will increase variance to $c^2 * variance$

Standard Deviation

- Given $s_x = \sqrt{Variance}$
- Tells us absolute spread of the data
- Adding a constant c does not change the standard deviation
- Multiplying by a constant results in standard deviation increaseing to $c * s_x$

Coefficient of Variation

- Given $\frac{s_x}{\bar{x}}$
- Quantifies degree of spread relative to the mean.

Median

- Middle value of the variable after arranging dataset in ascending/descending.
- If median is "in-between", divide by 2
- Adding a constant value to datapoints will increase median by that constant
- Multiplying a c to all datapoints will increase median to c*median

Interquartile Range

- 1st quartile, Q_1 is the 25th percentile data point
- 3rd quartile, Q_3 is the 75th percentile data point
- The interquartile range is given: $Q_3 - Q_1$
- Adding c to datapoints will not change the IQR
- Multiplying c to datapoints will increase IQR to $|c|$*IQR
- Interquartile range can't be negative.

Mode

- "Peak" of the distribution - the value that appears most frequently
- You can guess how this changes

Covariance and Correlation

- Covariance scales by c if all data points in one axis is multiplied by c
- Covariance scales by c^2 if all data points in both axes is multiplied by c
- Correlation, which is $\frac{Covariance}{std(X)*std(Y)}$, does not scale.

1.4 Study Designs

1.4.1 Experimental Studies

Experimental studies intentionally manipulate one variable in order to cause an effect on another. It established cause-effect relationship between 2 variables.
Experimental studies can provide evidence of cause-and-effect relationship

Treatment Group
The group in which the independent variable is manipulated to check for cause-effect

Control Group
The group which acts as the baseline for comparison - nothing is meddled with

Random Assignment (Randomised Experiment)
To control for other factors affecting results, random assignment should be used to assign participants into treatment/control books
If the groups are large, should result in similar groups, even though they might have different sizes. Note that it's still random if the chances of being in separate groups are unequal

Placebo
Treatment groups where the independent variable is manipulated but not in any meaningful way. Another baseline.
Note that the placebo can be a treatment with a known impact as well.

Blinding
Blinded subjects s.t. they do not know whether they are in treatment or control groups.
Can be done to assessors as well.

Double-Blind
If both subjects and assessors are blinded, the test is called a double-blind test.

1.4.2 Observational Studies

Observes individuals and measures variables of interest. Researches do not directly manipulate one variable to cause an effect.
Observational studies cannot provide evidence of a cause-and-effect relationship

In observational studies, we cannot ensure that confounders are controlled for

Treatment & Control
Even though the variables are not directly being manipulated, there still can be treatment/control groups.
However,they are assigned by the subjects themselves

2 Chapter 2

2.1 Understanding Rates

2.1.1 Rates

Let X be a category we're interested in. The rate(X) = $\frac{|X|}{n}$, where n is the total count in the population.

Joint Rate
For joint rates, we're look at something like what is the rate of X **AND** Y?
To get joint rate, $\frac{|X \cap Y|}{n}$

Conditional Rates
For conditional rates, we look at a statement like what is the rate of X **GIVEN** Y?
To get conditional rate, $\frac{|X \cap Y|}{|Y|}$

2.1.2 Association in Rates

When association is present:

- rate$(A|B) \neq$ rate$(A|\bar{B})$
- rate$(A|B) >$ rate$(A|\bar{B})$
    1. Presence of A when B is present is stronger than when B is absent

2. **Positive Association** between A and B
3. Equivalent to saying **negative association** between A and $\bar{B}$:
4. $\text{rate}(\bar{B}|A) < \text{rate}(\bar{B}|\bar{A})$

- $\text{rate}(A|B) < \text{rate}(A|\bar{B})$
  1. Presence of A when B is present is weaker than when B is absent
  2. **Negative Association** between A and B
  3. Equivalent to saying **positive association** between A and $\bar{B}$:
  4. $\text{rate}(\bar{B}|A) > \text{rate}(\bar{B}|\bar{A})$

When $\text{rate}(A|B) = \text{rate}(A|\bar{B})$, association is **NOT** present.

Symmetry Rule

$$\text{rate}(A \mid B) > \text{rate}(A \mid NB) \iff \text{rate}(B \mid A) > \text{rate}(B \mid NA)$$

$$\text{rate}(A \mid B) < \text{rate}(A \mid NB) \iff \text{rate}(B \mid A) < \text{rate}(B \mid NA)$$

$$\text{rate}(A \mid B) = \text{rate}(A \mid NB) \iff \text{rate}(B \mid A) = \text{rate}(B \mid NA)$$

2.1.3 Basic Rule on Rates

Overall rate(A) will lie between $\text{rate}(A|B)$ and $\text{rate}(A|\bar{B})$
As such, the closer rate(B) is to 100%, the closer rate(B) is to $\text{rate}(A|B)$

Additionally, if rate(B) = 50%, then rate(A) = $\frac{rate(A|B)+rate(A|\bar{B})}{2}$

Lastly, if $\text{rate}(A|B) = \text{rate}(A|\bar{B})$, then rate(A) = $\text{rate}(A|B) = \text{rate}(A|\bar{B})$

2.1.4 Bar Plots

Bar plots are good at visualizing categorical values and rates. visualize them now.

2.2 Simpson's Paradox

Simpson's paradox occuers when a trend appears in more than half of the groups of data, but **disappears/reverses** when the groups are combined

"Disappears" means that the two variables in question, (e.g. A & B) are no longer associated ($\text{rate}(A|B) = \text{rate}(A|\bar{B})$)

e.g. For Group A and Group B individually, treatment X is better. But when looking overall, treatment Y is better. OR no treatement is clearly better.

2.2.1 Simpson's Paradox Questions

1. Identify the main subgroups. It will be quite obvious like male and female. This subgroup will be **combined** to check for simpson's paradox
2. Identify the rate that might be reserved. This will be the most pertinent rate, like rate(S-grade—class)
3. Calculate the rate in the subgroups. It's better to think of it like rate(S-grade—class) than rate(S-grade—class and female), just ignore the subgroup for now.
4. Compare the rates in the two subgroups. Don't compare the subgroups, compare association between (in this e.g.) grade and class. Is grade still positively associated with this class?
5. Combine the subgroups. Is grade still associated with a certain class?

Explaining the cause of Simpson's Paradox

	Large stones			Small stones		
	Succ.	**Total trt.**	**R(succ.) in %**	**Succ.**	**Total trt.**	**R(succ.) in %**
X	381	526	72.4%	161	174	92.5%
Y	55	80	68.8%	234	270	86.7%
Total	542	700	77.4%	289	350	82.6%

From the table, comparing the number of treatments using X, we can see that majority of large-stones treatments use X, while comparatively the number of small-stone treatments using X is lower. By the basic rule of rates, the overall sucess of X will be closer to rate(Sucess X | Large Stones)

Conversely, treatement Y has higher number of treatments for small stones compared to large, so overall success of treatment will be closer to rate(success Y | Small Stone).

Therefore, overall Y seems to be the better option because the treatment succeeds for easier cases. But actually, X is better.
In this case, stone size is a confounding variable

2.3 Confounders

A confounder is a third variable that is associated with both the independent and dependent variable. It can be positively or negatively associated with the variables.

2.3.1 Checking for Confounders

Simply,

1. Check that the confounder (*kidney stone size*) is related to independent variables (*treatment type*)
2. Check that the confounder is related to the dependent variable (*treatment outcome*)
3. If so, then you've got a confounder.

2.3.2 Dealing with confounders

Randomised Assigment
Randomised assignment is a method to deal with confounding variables. Essentially, confounders occur due to association, which is a consequence of having unequal proportion of variables in two groups.

Doing random sampling (*to treatement type*), this would eliminate the confounder because it would no longer be associated with the independent variable.

3 Chapter 3

We'll look at a HDB dataset and our question is "What factors may affect the pricing of resale flats sold in SG?"

3.1 One-Variable Exploratory Data Analysis (EDA)

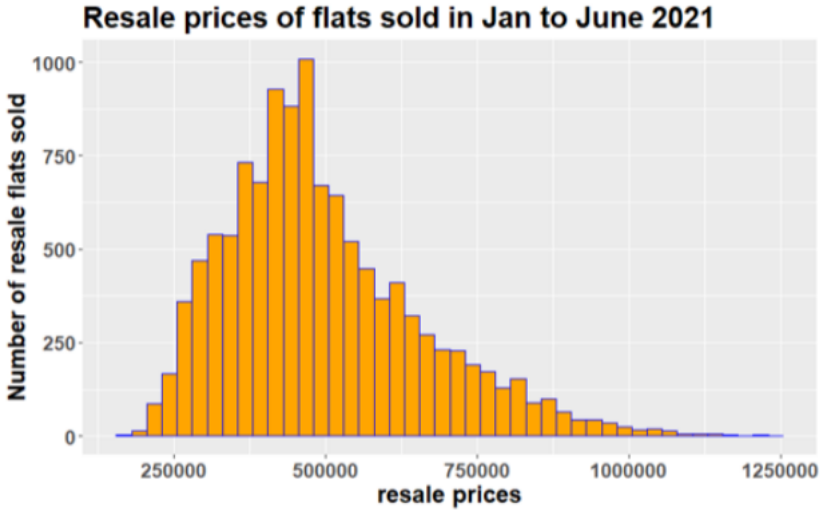
To investigate the distribution of some variable (*age*), we create a frequency table showing how much observations exist for each variable value:

Age	**Frequency**
2	9
3	8
4	583
5	1105
6	884
7	295
8	255
⋮	⋮

We now look at how 2 graphs can help us visualise this:

3.1.1 Histograms for Distribution

Histograms group values into equal-sized bins (*e.g. 2-4 year old HDBs*):

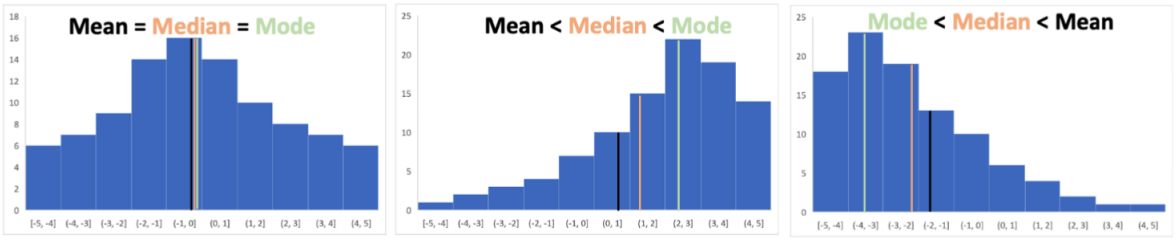


We describe histograms with their shape, center, and spread.

Shape - Peaks and Skewness
In the histogram, we can see a distinct peak at [455k,480k]. This histogram is unimodal, but if there were two peaks it'd be bimodal

As for skewness, there is a long tail **pointing to the right**, so it's right-skewed. If the tail was pointing left, it'd be left-skewed. If there were tails towards the left and right, it'd be symmetrical.

Central Tendency - mean, median, mode
We can also describe histograms using values.



Looking at the mean, median and mode and how they compare to each other usually tells us how it skews.

Spread - Standard Deviation and Range
Lower standard deviation, gives a more defined peak. Additionally, range (which can be misleading) tells us something about variation as well.

AVOID!
Avoid using bins that are too large or small. Experiment with bin width.

3.1.2 Outliers

Outliers are data points such that:

- The value of the data point is **greater** than $Q_3 + 1.5 \times IQR$
- The vlaue of the data point is **lesser** than $Q_1 - 1.5 \times IQR$

Generally, we do not want to remove outlier data points from a dataset because they tell us something about the behavious of the data. Instead, see why such extreme values happen

3.1.3 Boxplots for 5numsummary

The boxplot easily visualizes the five number summary of the dataset, which is:

1. Minimum
2. Q1
3. Median
4. Q3
5. Maximum

To create a boxplot,

1. Draw a box from Q_1 to Q_3.
2. Draw a line in the box where the median is
3. Identify outliers
4. Extend a line from Q_1 to the smallest value that is not an outlier
5. Extend a line from Q_3 to the largest value that is not an outlier
6. Mark each outlier with a dot

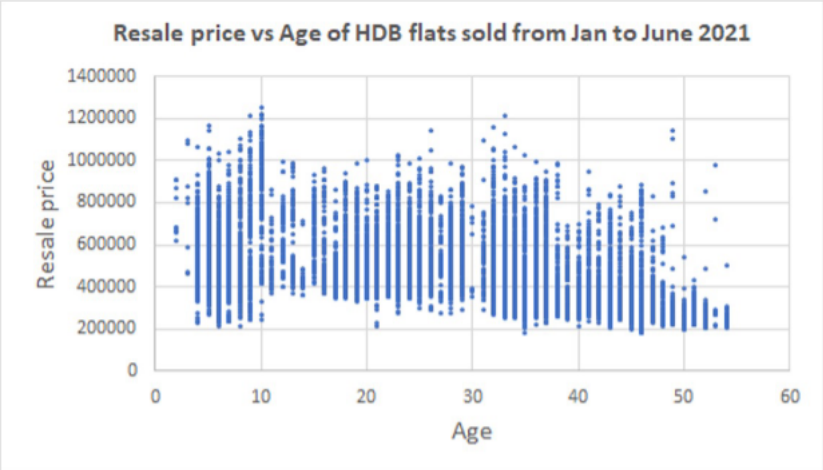
3.2 Two-Variable EDA

A relationship between two variables is *deterministic* if the value of one variable can be determined exactly if we know the value of the other variable. (e.g. cm to m)

We look at *statistical* or non-deterministic relationships in this section, where given one variable we can describe the average value of the other variable

To visualise two variables, we plot scatterplots.

3.2.1 Scatterplots



The scatterplot above shows the independent value, age, on the x-axis. The **dependent variable**, resale price, is on the **y-axis**.

We use direction, form, and strength to describe scatterplots.

Direction

1. Positive Direction/Relationship
   (a) Graph looks like an increasing slope, y=x
   (b) Increase in one variable leads to increase in the other
2. Negative Direction/Relationship
   (a) Graph looks like decreasing slope, y=-x
   (b) Increase in ove variable results in decrease in other
3. Neither Positive nor Negative
   (a) Graph has no clear direction in one way or the other
   (b) Behaviour is erratic or nonlinear

Form
Form is the general shape of the scatter plot.

It is linear when the data points *appear* to be scattered about a straight line.

It is non-linear when the data points appear to scatter about a smooth curve

Strength
Strength indicates how closely the data follow the form of the relationship.
Are they scattered about loosely? Or tightly follow a line?

3.3 Correlation Coefficient

Correlation coefficient tells us the measure of **linear** association between two variables.
It ranges from -1 to 1, which summarizes both the direction and the strength of the assocaition between the variables.

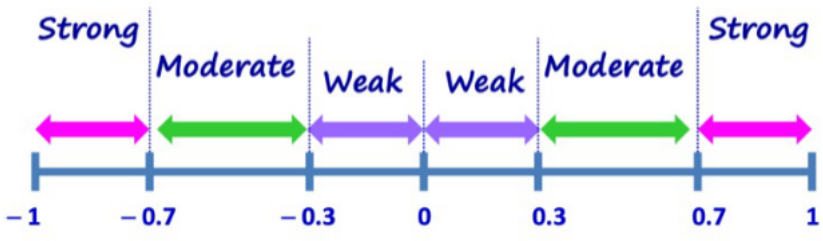
A negative value indicates a negative relationship, while positive the opposite.

Note that if the correlation coefficient is 0, there is no linear association between the two variables.
A non-linear relation **might** exist, or there could be no relation at all.

3.3.1 Correlation coefficient calculation

Standard units are calculated, given:

$$\frac{x-\bar{x}}{std(x)}$$



Additionally, a correlation coefficient of 1/-1 is a **perfect** positive/negative correlation.

3.3.2 Correlation coefficient does not change based on axis

No matter which axes the dependent/independent variable is plotted on, the correlation coefficient does not change.
Additionaly, correlation coefficient does not change if a constant is added to OR multiplied to data points.

3.3.3 Notes on correlation!

1. **Association is not causation** A strong linear assocation does not mean that a change in variable x will cause a change in variable y. Establishing linear association is simply establishing a *statistical relationship*, not a causal one.

2. Correlation does not tell us anything about non-linear association. Other associations such as logarithmic or cubic ones could still exist.

3. Outliers can affect coefficient significantly. It's possible they don't affect anything as well.

3.3.4 Ecological Correlations

Corrleations that occur in sub-group levels (Individual level), may differ dratically from correlations made taking aggregate (Ecological).

Individual level correlations observed do not hold for the aggregate (**Ecological**) level, and vice-versa.

This links to fallacies discussed later.

3.4 Linear Regression

If 2 variables are linearly associated, we may model the relationship by fitting a straight line to the observed data. This is called linear regression.

Using the previous example, if we fit a linear regression line on Age vs Resale price, we get something like:

$$y = -4007x + 591587$$

Where y is the resale price (dependent) and x the age (independent).

We may use any arbitrary x value to **predict** how the flat would sell at age x. Note that this is only a prediction

3.4.1 Least Squares

This is one of the methods to get a regression line. It minimises the sum of squares of errors.

3.4.2 Notes on regression lines

1. They always pass through the average value in the dataset: $(\bar{x}, \bar{y})$

2. It does not allow for prediction of x value given y

3. The correlation coefficient is used in calculating the gradient of the line: $\frac{s_Y}{s_X} \times r$, where s is the standard deviation and r the correl

4. Should only be used to predict values that lie **inside** the range it was trained on.

3.4.3 Modelling non-linear associations

If a relationship between to variables x and y are a known curve, for example exponential, we can manipulate them to obtain linear associations.

If a relationship is exponential, we can instead plot ln y against t: $y = cb^x \rightarrow ln\ y = ln\ c + x\ ln\ b$
Steps:

1. For each data point x, y, compute (x, lny)
2. Plot ln y against t
3. Find a linear regression for ln y vs t
4. Using the regression line, find values for c and b
5. Sub it into the original equation
6. You now have an exponential relationship for x and y

3.4.4 Fallacies

Fallacy	Using	To conclude
Ecological	Ecological correlation (aggregate level)	Individual level correlation
Atomistic	Individual level correlation	Ecological correlation (aggregate level)

4 Statistical Inference

4.1 Probability

Uniform probability is the way of assigning probabilities to outcomes such that equal probability is assigned to every outcome in the finite sample space.

4.1.1 Conditional Probability and Independence

Written as P(E|F), which is the probability of E *given* F.

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

We restrict the sample space to F, and count occurences where E and F both occur.
Note that probabilities are equivalent to rate:

Random sampling	Corresponds to	Probability experiment
Random Sampling	Corresponds to	Probability experiment
Sampling frame	Corresponds to	Sample space
A subgroup A of the sampling frame	Corresponds to	An event A of the sample space
The rate of A, rate(A)	Corresponds to	The probability of A, $P(A)$
Conditional rate rate(A|B)	Corresponds to	P(A|B)

4.1.2 Joint Probability

$$\text{P(A and B)} = \text{P(A}\cap\text{B)} = \text{P(A)} + \text{P(B)} - P(A \cup B)$$

4.1.3 Prosecutor's Fallacy

Prosecutor's fallcy is erroneously taking the P(A|B) to be equal to P(B|A). This is only truen when P(A) = P(B) or P(A∩B)=0.

4.1.4 Independence of Events

If two events A and B are independent then

$$P(A) = P(A|B)$$

Implying that $P(A) \times P(B) = P(A \cap B)$

It is equivalent to stating that the two variables are not associated

Conditional Independence
We can extend this to conditional probability. Two events A and B are conditionally independent if:

$$P(A \cap B|C) = P(A|C) \times P(B|C)$$

4.2 Fallacies and Random Variables

4.2.1 Law of Total Probability

The law to total probability states that if E, F, G are events from some sample space S s.t.

1. E and F are mutually exclusive
2. $E \cup F = S$

Then,

$$P(G) = P(G|E) \times P(E) + P(G|F) \times P(F)$$

4.2.2 Conjunction Fallacy

Occurs when one believes that the chances of **two** things happening together is higher then the chance of **one** of those things happening alone.

$$P(A \cap B) \leq P(A) \quad \text{and} \quad P(A \cap B) \leq P(B)$$

4.2.3 Base Rate Fallacy

The base rate fallacy is a decision making error in which information of the rate of occurence of some trait in a populatio (base rate), is ignored or not given appropriate rate.

Example
It is given that for some breathalyzer test:

1. 5% chance of false positive; that sober drivers test positive
2. 100% chance of true positive; that drunk drivers test positive

If we only use this information to conclude that it is a good test, we commit base rate fallacy.

Now consider that only 1 in 1000 drivers drive drunk. What is the probability of (Drunk|Positive Test)?
Now, we can see that testing positive is actually much more associated to sober drivers than drunk drivers. This is a bad test.

	Positive test	Negative test	Total
Drunk driver	1	0	1
Sober driver	49.95	949.05	999
Total	50.95	949.05	1000

Sensitivity
The sensitivity of some medical test is the *true positive* rate:

$$\text{P(Test Positive} \mid \text{Infected)}$$

Specificity
The specificity of some medical test is the *true negatvie* rate:

$$\text{P(Test negative} \mid \text{Not Infected)}$$

4.2.4 Random variables

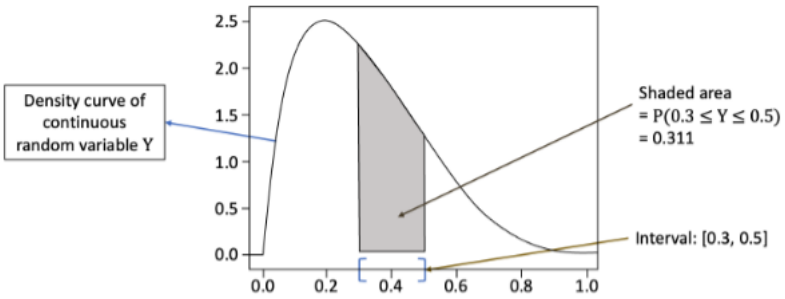
Ranom variables are numerical variables with probabilities assigned to each possible numerical value.
e.g. For rolling a dice the random variable of the value 3 is:

$$\text{P(Y} = 3) = \frac{1}{6}$$

For dice rolls, they are **discrete random variables** as dice roll values are discrete.

Continuous Random Variables
For variables that take on any continuous number, continuous random variables are used.

Continuous Random variables are not defined at specific values. Instead, they are defined by the area over an interval of values.



4.3 Statistical Inference and Confidence Intervals

4.3.1 Statistical Inference

Statistical Inference refers to the use of samples to draw inferences or conclusions about the population in question.
However, using a sample to estimate the population parameter is subject to inaccuracies:

$$\text{Sample statistic} = \text{population parameter} + \text{bias} + \text{random error}$$

To make inference about the population, the fundamental rule for using data for inference should be met.

Fundamental Rule for using Data for Inference
Available data can be used to make inferences about a much larger group if the data can be considered to be representative with regards to the question of interest.
This can be done by reducing selection bias (e.g. using a good sampling frame and Simple Random Sampling) and non-response bias. If there is little bias, then:

$$\text{Sample statistic} = \text{population parameter} + \text{random error}$$

Random error is the small differences that aries as a result of the sampling variability.

4.3.2 Confidence Intervals

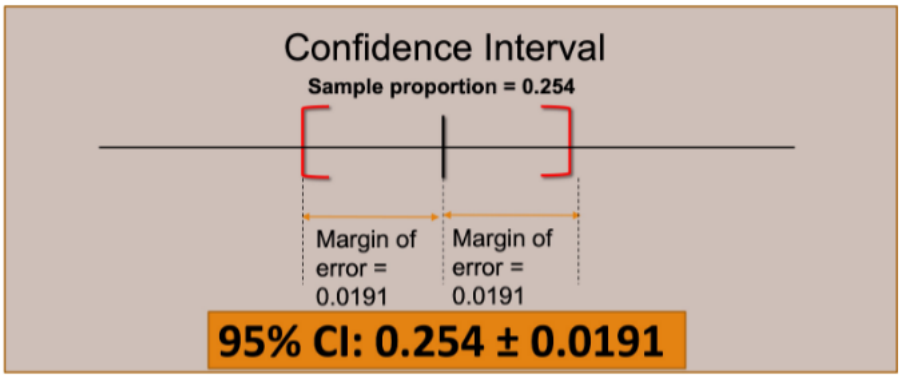
Confidence interval is a range of values that is likely to contain a population parameter based on a certain degree of confidence (**confidence level**). It is given:

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

Where p^* is the sample proportion, z^* is the "z-value", and n is the sample size.

Z-value is based on our confidence interval. For 95% confidence, it is 1.645

After computation, the CI value will look something like: $\mathbf{x}\pm\mathbf{y}$, where x is the sample statistic and y is the variation:



This tells us:

We are 95% confident that the population proportion of [dependent variable] that are [independent variable] lies within this confidence interval

4.3.3 Confidence levels

What 95% means in the above context, is that if many simple random samples of the same size are taken, and a CI is constructed for each of them, about 95$ of CIs constructed contain the population parameter.
Thus, looking at the above statement it is wrong to say:

There is a 95% chance that the population proportion of 5-room resale flats lie between 0.235 and 0.273

Because,

1. The population proportion is unkown to us
2. For any particular sample, the confidence interval constructed only depends on the sample proportion and the value of z^* corresponding to a chosen confidence level. Thus, the confidence is also fixed and there is no probabilistic element.

It instead tells us we are 95% **confident** that the population statistic/proportion lies within the given confidence interval

4.3.4 Properties of Confidence Intervals

1. The larger the sample size, the smaller the random error. This results in a narrower confidence interval
2. The higher the confidence level, the wider the confidence level

4.4 Hypothesis Testing

4.4.1 Hypothesis Test

A hypothesis test is a statistical inference method used to decide if the data from a random sample is sufficient to support a particular hypothesis about a population.
It asks if our observed sample proportion's deviation from the hypothesised proportion can be explained by chance variation.

Steps of Hypothesis testing for population proportion

1. Identify the question and state the null hypothesis and alternative hypothesis.
   (a) The **null hypothesis** corresponds to the case where our observation can be explained by chance variation
   (b) The **alternate hypothesis** corresponds to the case where our observation is not due to random chance.
2. Set the significance level of the test. (1%/5%/10% are common)
3. Find relevant sample statistic from sample
4. Calculate p-value
5. Make conclusion of hypothesis test
   (a) p-value less than significance level: reject null in favour of alternative
   (b) p-value more than equal to significance level: Inconclusive test

4.4.2 P-values

The p-value is the probability of obtaining a result as extreme or more extreme than our observation, in the direction of the alternative hypothesis, **assuming the null hypothesis is true**.

4.4.3 Hypothesis test for population mean

The t-test is used to perform the hypothesis test for population mean.
The steps are exactly the same tho

4.4.4 Hypothesis test for association

The chi-squared test is done to test for association between categorical variables in a population. Steps to do so are similar, but slightly different:

1. Identify question and state null and alternative hypothesis:
   (a) Generally, the research question would be "is there an association between x and y"
   (b) Hence, **null hypothesis** would be something like: There is no assocation between x and y
   (c) **Alternative hypothesis** would be something like: There is an association between x and y
2. Assume null hypothesis
3. We'd expect that using the rates of X we can find out values of $\bar{X}$ accurately, as there is no association
4. Calculate p value and conclude.

Note that the Chi-squared test only tells us about the existence of association, not strength or direction.

4.5 Rejecting hypothesis tests

If we fail to reject the null hypothesis, and the test is inconclusive, we say that we have insufficient evidence to support the alternative.