ACU
AUSTRALIAN CATHOLIC UNIVERSITY

PUBH620 Biostatistics

Life in Statistics

Week 2

Inferential Statistics and Sampling Techniques

Lecture notes by Dr Brandon Cheong and Dr Peter Mahoney

ELECTRONIC WARNING NOTICE

In this lecture…

(clicking the links below will direct you to the topic page)

Topic Learning Objectives (TLOs)

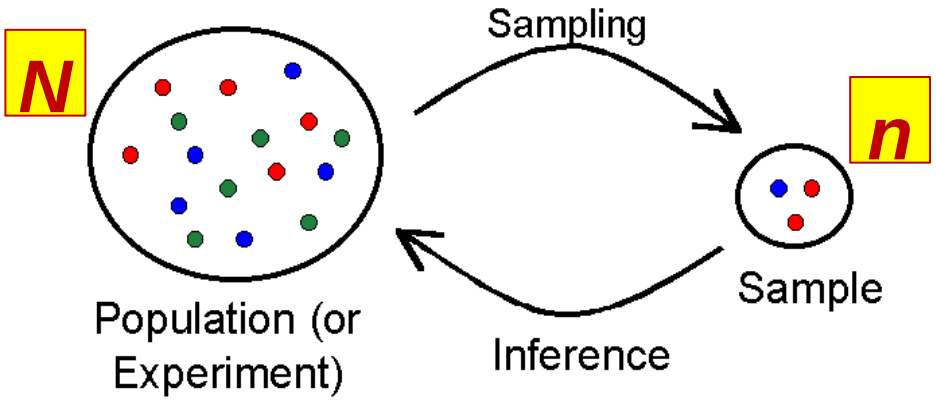
- Understand the need for inferential statistics in public health
- Understand standard error and confidence intervals
- Understand the concept of the p-value and the role it plays in statistics
- Distinguish between non-random and random sampling techniques in public health
- Distinguish between Type I and Type II errors

2.1 Inferential Statistics

Inferential statistics

Previously we mentioned that it is often impossible to measure or analyse an entire population of interest. Research studies in health are usually made on samples of subjects (patients, participants etc.) and not whole populations.

The key challenge here is drawing a sample that is representative of the population. This is what we call **inferential statistics.** By taking a sample, we are drawing an inference about its population.



2.1 Inferential Statistics

Inferential statistics

However, there are a few things to consider:

- Standard error, which is the error associated with sampling – we understand that the means and standard deviations of the population and samples are not going to be exactly the same.
- Confidence Intervals – how confident are we that the sample is representative of the population?
- Sampling method – What is the best way to remove/reduce sampling bias?
- Assessing the degree of bias – to assess the degree of bias, we must first understand the population from which the sample was drawn.
- In practice – very difficult to completely eliminate some form of sampling bias, which we then determine to what extent should it be present?

2.1 Inferential Statistics

Sample

A sample is some fragment of the whole population. In practice, clear inclusion and exclusion criteria need to be stated to show how the sample was obtained.

Samples must be representative. The best way to do this is to ensure that each participant/patient has an equal chance of being selected as part of the study. This is often done through random sampling.

Population

A population includes everyone usually with a particular set of characteristics (same disease, geographical location, sex, socioeconomic status, time of diagnosis etc.)

These characteristics help to define the study population.

2.1 Inferential Statistics

How to conduct a study

1. Define your study population and draw a sample from it
2. Define study parameters (age, sex, occupation, geography etc.)
    - You can base these on states (VIC, NSW etc.), local government authorities (LGAs), healthcare administrative boundaries (eg. Medicare locals and hospital networks)
3. Consider how representative your sample is of the population.
    - Eg. Census and LGA data is usually precise.
    - Questions to consider: Could large Australian states be representative of all Australia? Or could LGAs be representative of all Australia?
4. Collect your required data

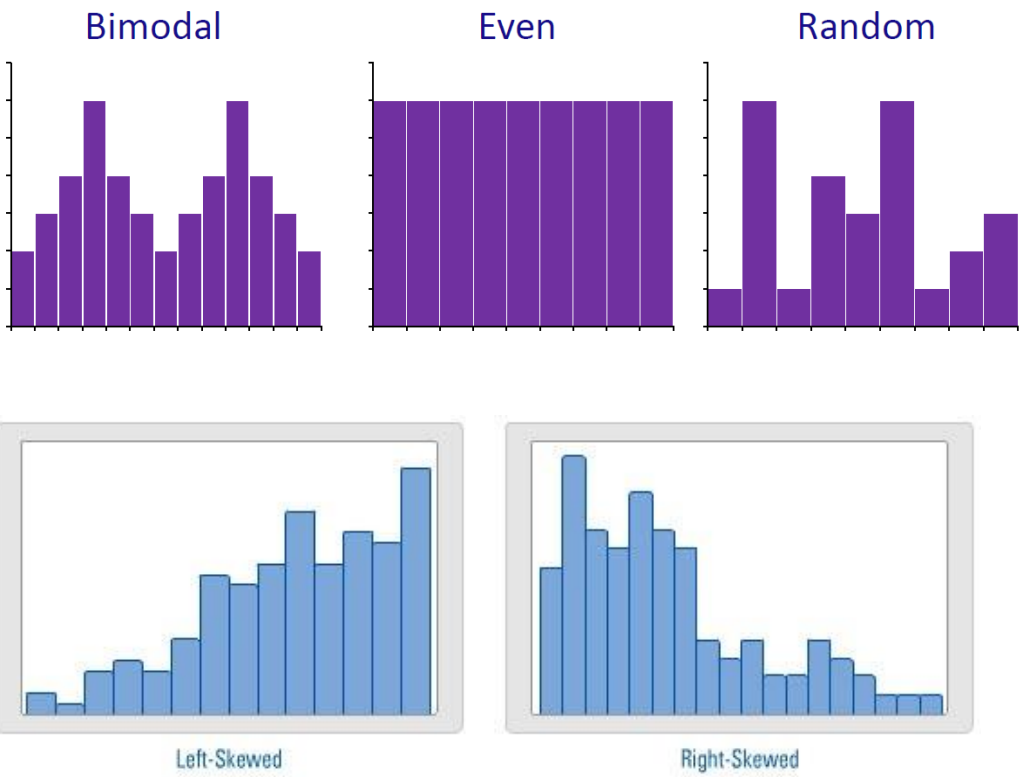
2.1 Inferential Statistics

Why do we need inferential statistics?

- Usually very difficult to study an entire population (however, some hospital databases may allow for this).
- Descriptive statistics is great for getting central values and descriptive summaries of our data but it doesn’t help us draw conclusions about our study.
- Inferential statistics allows us to do a study that goes beyond our immediate data. We can infer what a population mean might be by calculating a sample mean.
- Very useful in health research for making predictions eg. Effectiveness of a new cancer drug.
- It is central to hypothesis testing.

2.2 Normality and Tails

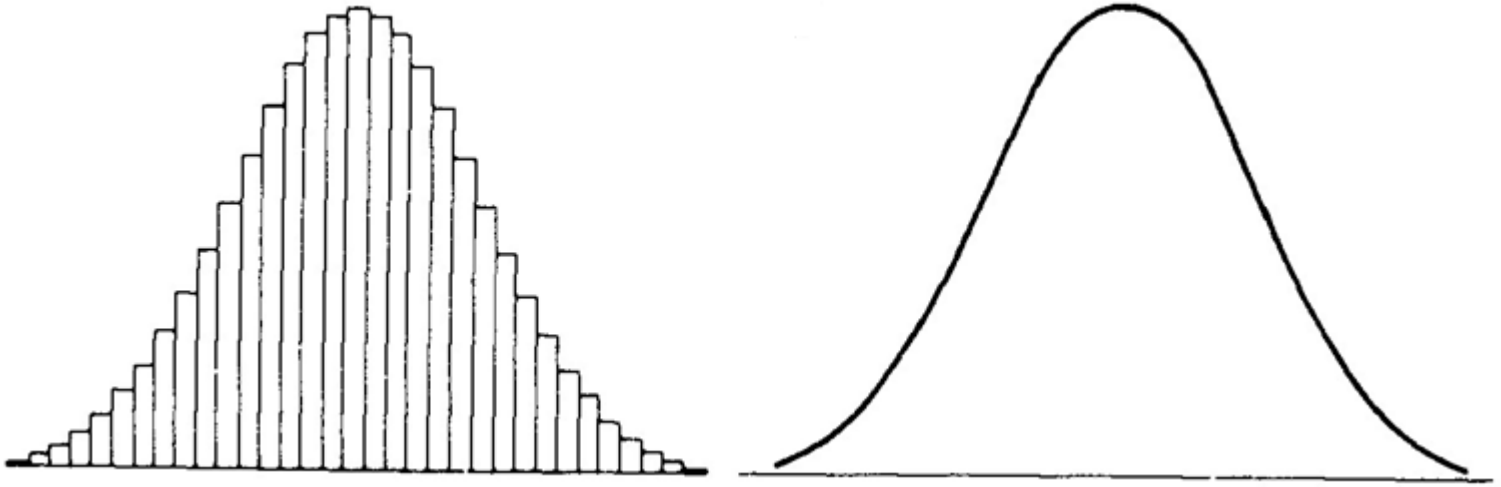
Distribution is also important in hypothesis testing. Common distributions are:



2.2 Normality and Tails

One of the most important distributions in statistics is the normal curve:

- A bell-shaped curve that is symmetrical around the centre.
- Smooth line often used with continuous variables.

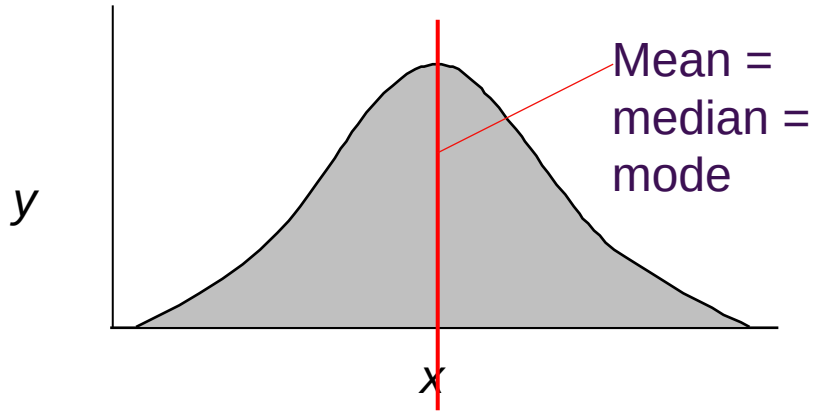


2.2 Normality and Tails

The normal distribution is important in statistics because many (parametric) tests often assume that the data follow a normal distribution. Such tests include two sample t-tests, ANOVA and Pearson's correlation test.

Properties of a normal distribution are:

1. Symmetrical around population mean (μ)
2. Mean = median = mode
3. Spread described by population Std. Dev. (σ)
4. Usually μ and σ estimated from $\bar{x}$ and s

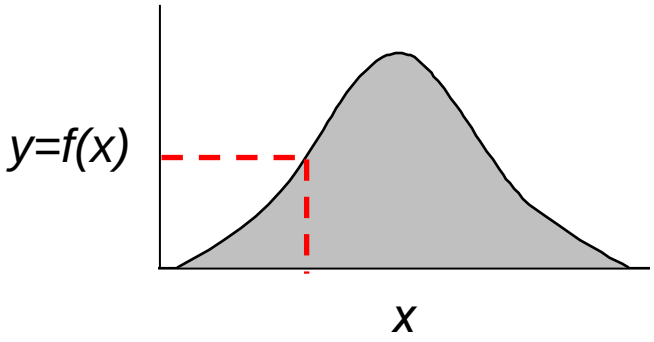


2.2 Normality and Tails

Mathematics of the normal distribution

The normal distribution curve is a probability function where y = f(x) and the area under the curve is total probability = 1 (100% of observation are within the curve).

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



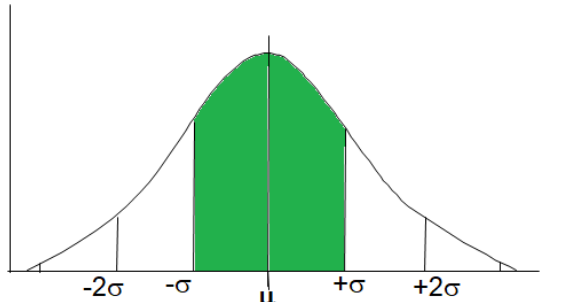
***e* = constant**

π (Pi) = constant

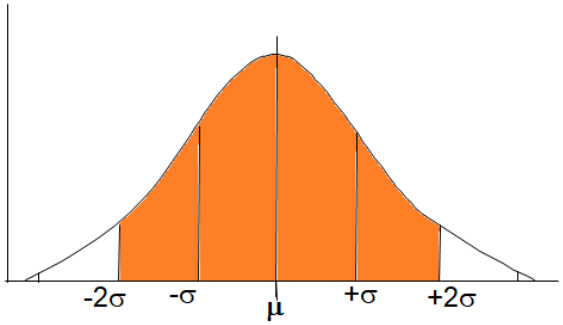
Note: you do not need to know how to use this formula but only understand the variables in it.

2.2 Normality and Tails

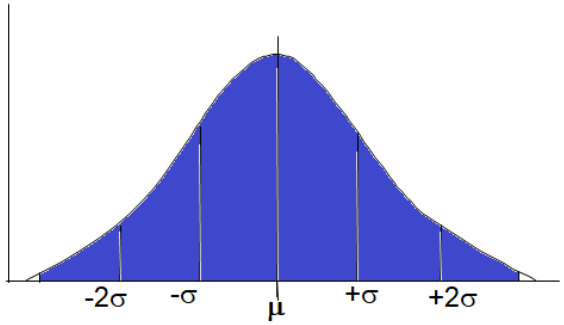
Mathematics of the normal distribution



68% of observations within 1 SD of mean



95% of observations within 2 SD of mean



99% of observations within 3 SD of mean

2.2 Normality and Tails

It often makes more sense to talk about probabilities of 0.68 (68%), 0.95 (95%) and 0.99 (99%).

The two most common probabilities that we see in statistics are 0.95 and 0.99. Often we might talk about a “95% confidence interval” or “99% CI” (more on this later).

It is generally more accurate to say…

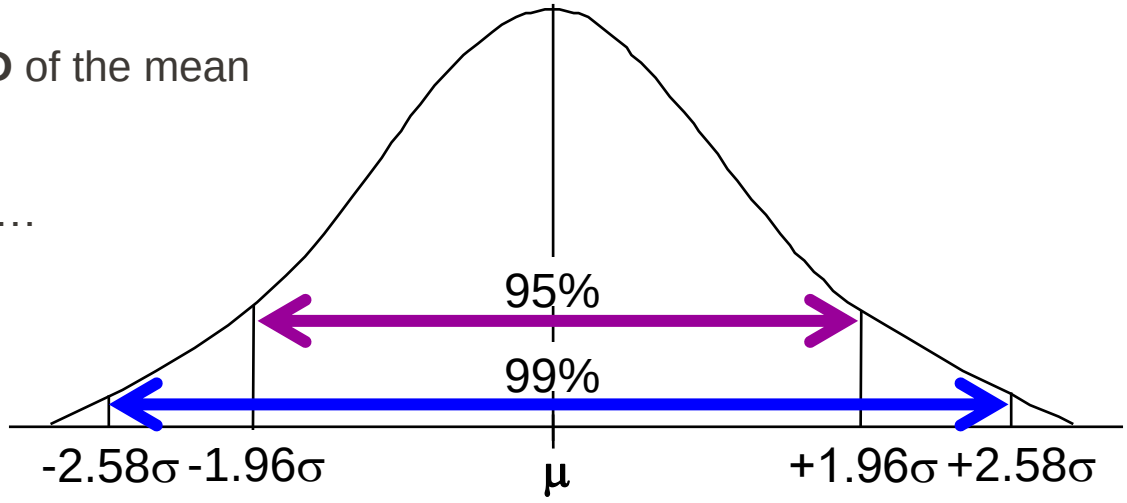
0.95 observations fall within **1.96 SD** of the mean

Or

0.99 observations fall within **2.58 SD** of the mean

This is due to the t distribution table…

In other words, the probability of choosing an observation outside of these limits is <0.05 and <0.01 respectively.



2.2 Normality and Tails

The t distribution table

By convention we use an $\alpha = 0.05$ and a two-tailed (two-sided) test. This is what is accepted universally.

α is known as the confidence level and helps us determine if we have statistical significance.

The degrees of freedom is n – 1, the sample size minus 1.

How do we determine the type of tailed test to use?

Note: In practice, SPSS is more efficient than using distribution tables but it is good to know the theory behind the statistics.

Degrees of Freedom	Alpha Level for one-tailed tests			
	0.05	0.025	0.01	0.005
	Alpha Level for two-tailed tests			
	0.1	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.269
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.043	2.462	2.756
30	1.697	2.042	2.457	2.750
∞	1.645	1.960	2.326	2.576

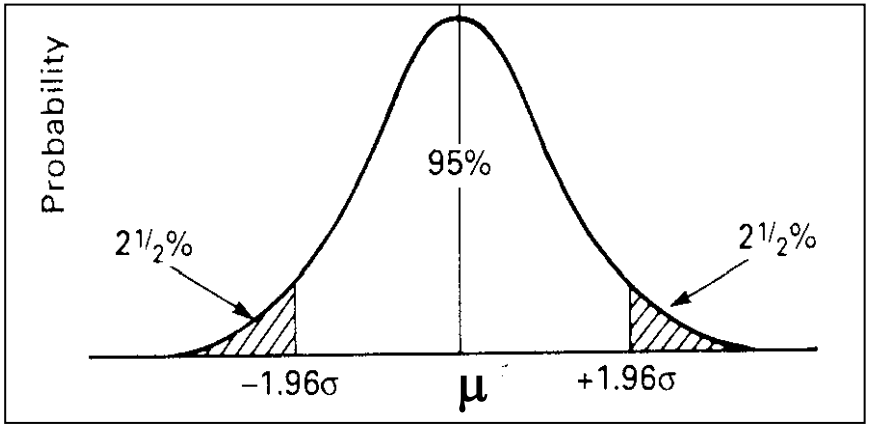
2.2 Normality and Tails

One and two-tailed testing

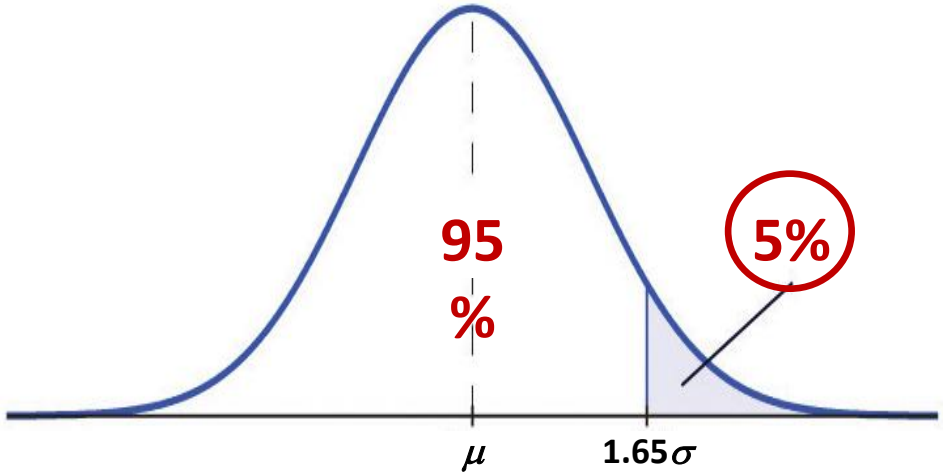
Hypotheses tests may be either directional or non-directional.

A two-tailed test allows for the possibility that differences between groups can occur in either direction.

A one-tailed test only allows for differences to occur in one direction.



Two-tailed test



One-tailed test

2.2 Normality and Tails

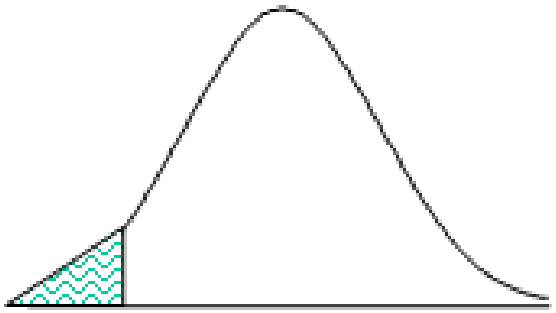
Example: Life span of humans living in Mozambique (Group A) and Sierra Leone (Group B).

If the mean life span of both groups were not different – a two-tailed test would be used because there is no directionality.

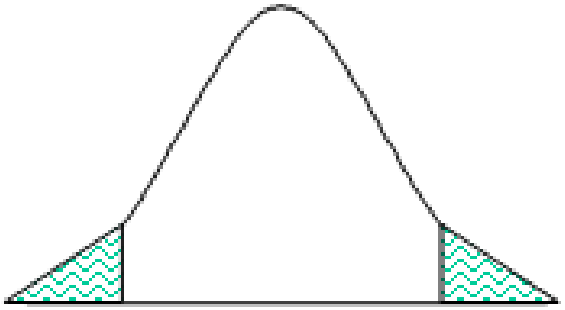
Suppose we thought the mean lifespan of Group A was longer than Group B – a positive one-tailed test would be used as we have said the lifespan of Group A > lifespan of Group B.

If we said Group A < Group B, then a negative one-tailed test would be used.

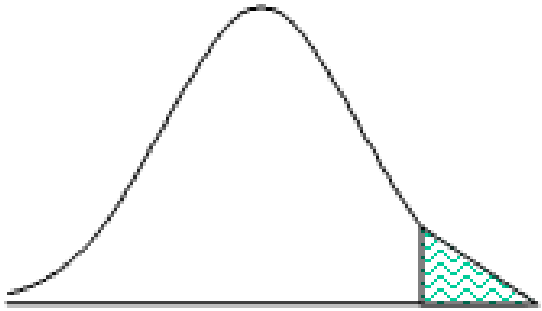
Good to know the above but conventionally, we use a two-tailed (2-sided) test.



Negative one-tailed test



Two-tailed test



Positive one-tailed test

2.3 Standard Error

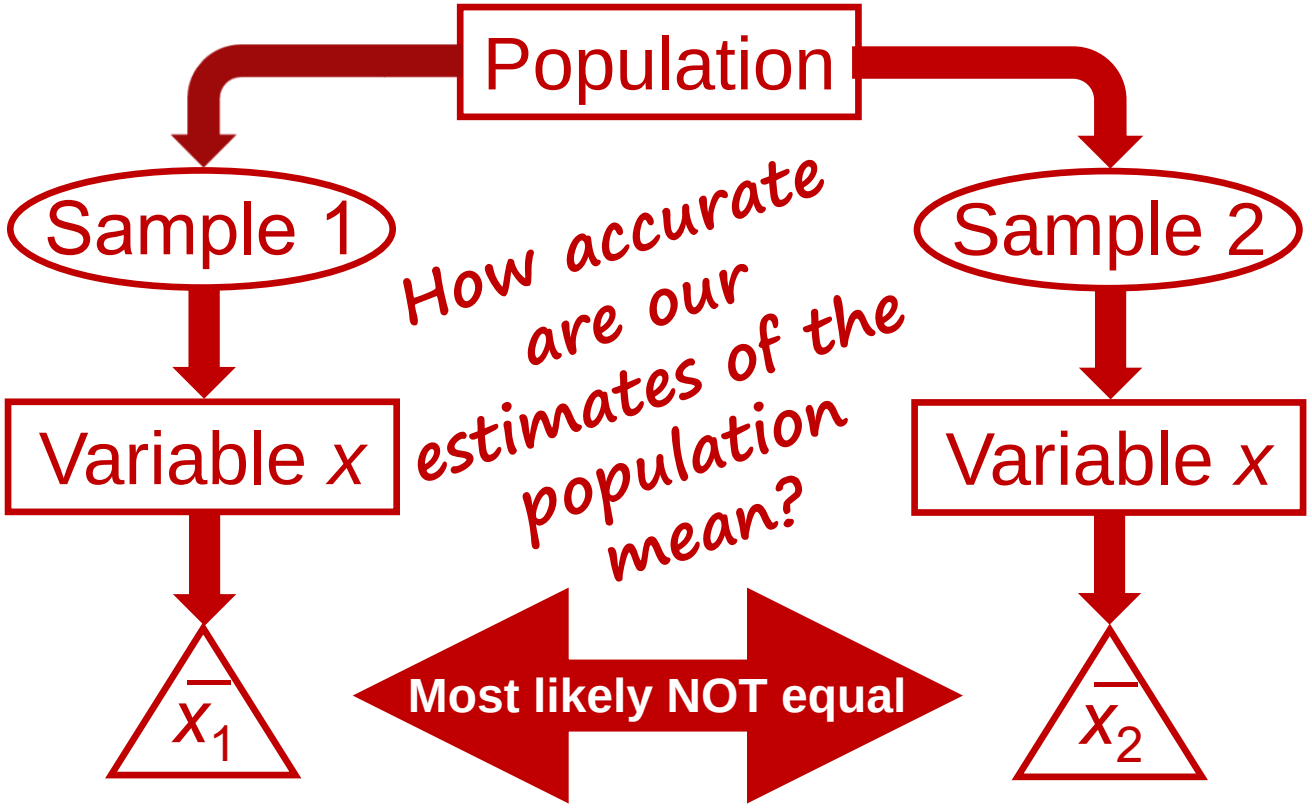
Standard error is the error associated with sampling. We can calculate sample mean ($\bar{x}$) and std. dev. (s) but they will never be exactly the same as the population mean (μ) and std. dev. (σ), only estimates.

Each sample is different so we can never be 100% sure that our sample is representative.

Eg. Conducting a study on CEO cortisol levels in Melbourne CBD. Each sample we obtain may have different numbers of male to females, or different age groups etc. Hence we can see how sampling error is present. To determine the error, we calculate the **standard error of the mean**.

This brings us to the idea of the **central limit theorem**, which states that the means of many random samples of a population are normally distributed. This then suggests that the means of the sample means are the population mean.

2.3 Standard Error



2.3 Standard Error

Estimates error associated with sample means:

$$SE = \frac{s}{\sqrt{n}}$$

← **Sample standard deviation**

← **Sample size**

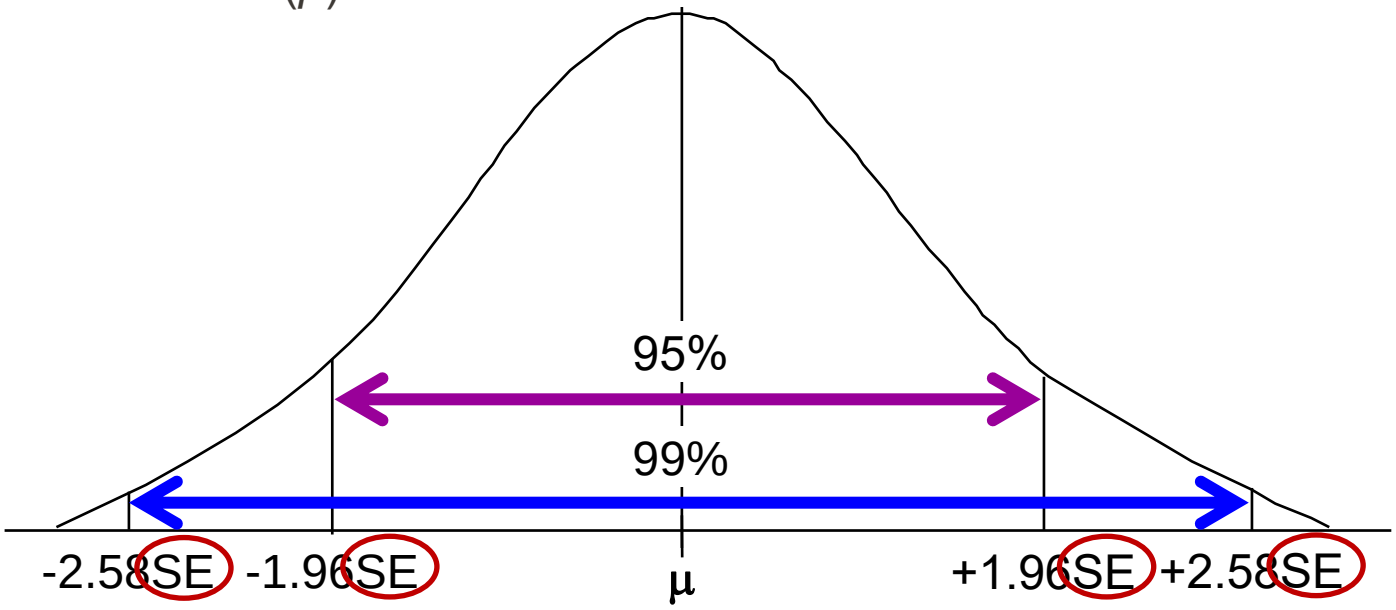
From the above formula, we can see that having a larger sample size n, will reduce SE. A smaller SE is better.

SE indicates how good an estimate our sample means ($\bar{x}$) are of our population mean (μ).

2.3 Standard Error

The properties of a normal distribution hold for a normal distribution of sample means.

Given a large sample size, 95% of our sample means ($\bar{x}$) are within ± 1.96SE of the population mean (μ).



Note: now SE, not SD

2.4 Confidence Intervals

Confidence Intervals

Are a range of values given as an estimate of where the population mean lies.

Eg. (From previous slide) “Given a large sample size, 95% of our sample means ($\bar{x}$) are within ± 1.96SE of the population mean (μ).”

We are saying that we are 95% confident that the population mean (μ) is somewhere within $\bar{x}$ ± 1.96SE . Hence this range is called the 95% CI (confidence interval)

Works the same for a 99% CI ie. $\bar{x}$ ± 2.58SE

How do we determine the upper and lower limits of our 95% or 99% CI?

2.4 Confidence Intervals

Example: Patient weights (kg) in a public hospital

n = 272 patients, s = 30.92, $\bar{x}$ = 99.40

t at a 95% CI (α = 0.05) = 1.96 (2-sided)

t at a 99% CI (α = 0.01) = 2.58 (2-sided)

Found from t distribution table

Formulas:

$$SE = \frac{s}{\sqrt{n}}$$

$$CI(95\%) = \bar{x} \pm t \times SE$$

2.4 Confidence Intervals

For a 95% CI…

$$SE = \frac{30.92}{\sqrt{272}} = 1.88$$

$$CI(95\%) = \bar{x} \pm t \times SE$$

Break this down:

$$Lower\ limit = 99.40 - 1.96 \times 1.88 = 95.71$$
$$Upper\ limit = 99.40 + 1.96 \times 1.88 = 103.08$$

We are 95% confident that the true value of the population mean is between 95.71 kg and 103.08 kg.

Patients were most likely overweight!

2.5 P-values and Statistical Significance

In week 1, we learnt that there are generally six steps to hypothesis testing. These are:

1. Develop a hypothesis (known as H_1)
2. State the null hypothesis (known as H_0)
3. Design an experiment to test H_0
4. Collect appropriate data
5. Calculate test statistic
6. Find the p-value and form a statistical conclusion

Steps 5 and 6 above are to do with analysing and presenting your data.

2.5 P-values and Statistical Significance

Analysing data

Generally involves a three-step process

1. Calculate the test statistic of your chosen test (a standardised test score)
2. Calculate the p-value (usually done through SPSS)
3. Compare p-value to α (your chosen confidence level)

There are a large variety of statistical tests available to us but most of them follow the three-step process.

This allows for different variables or groups to be compared.

It also allows for the examining of relationships of different variables and groups.

2.5 P-values and Statistical Significance

P-value

P is a measure of probability (but this is not the full story!).

We cannot simply say that P is the probability because a condition holds.

Remember the null hypothesis, H_0?

A p-value is defined as the probability of getting a result (whether standard or extreme) when the null hypothesis of the study question is true.

A convoluted explanation! In essence, it is a value that helps us determine whether there is any statistical significance in our results.

It helps us to realise that the results that we have from our data didn't just occur by chance!

2.5 P-values and Statistical Significance

Statistical significance

Given $\alpha = 0.05$ (by convention)

A p-value < 0.05 is said to be **statistically significant** and so the null hypothesis is rejected. We have evidence to support the alternate hypothesis.

A p-value > 0.05 is said to be **not statistically significant** and so we fail to reject the null hypothesis.

It is always important to report a p-value in your work! This will help you understand the significance of your results!

2.5 P-values and Statistical Significance

A common statistical test for checking normality

The Kolmogorov-Smirnov test (K-S test) is a test done in SPSS that checks that a dataset fits a normal distribution.

You have a dataset on patient heights (cm) with a sample size (n = 22).

Does this dataset follow a normal distribution? Develop your hypothesis.

H_0 = There is no difference in the distribution of this data set and that of a normal distribution

H_1 = There is a difference in the distribution of this data set and that of a normal distribution

Conduct a K-S test using SPSS! Try it yourself by importing the “patient height cm.xlsx” file into SPSS and running a 1 sample K-S test.

2.5 P-values and Statistical Significance

H_0 = There is no difference in the distribution of this dataset and that of a normal distribution

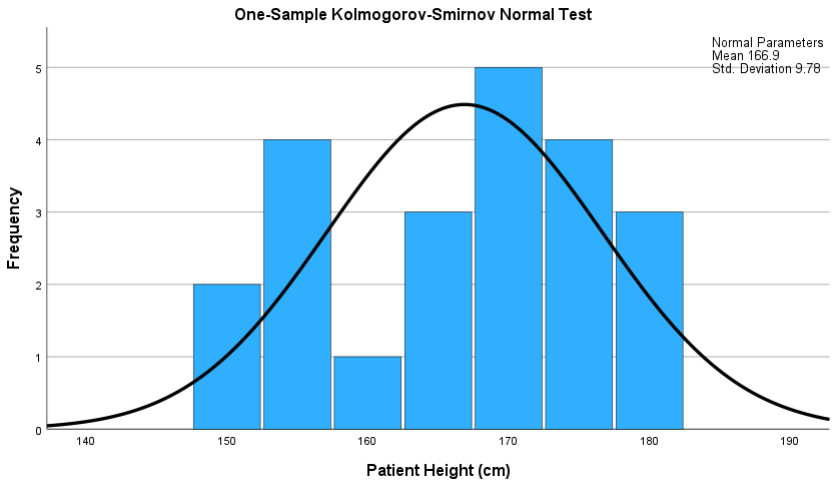
H_1 = There is a difference in the distribution of this dataset and that of a normal distribution

One-Sample Kolmogorov-Smirnov Normal Test Summary

Total N		22
Most Extreme Differences	Absolute	.132
	Positive	.116
	Negative	-.132
Test Statistic		.132
Asymptotic Sig.(2-sided test)[a]		.200[b]

a. Lilliefors Corrected

b. This is a lower bound of the true significance.



N = 22

Test statistic from SPSS analysis = 0.132

P-value (Asymptotic Sig. (2-sided test)) = 0.200

Here we see that p (0.200) > α (0.05) therefore we

accept the null hypothesis and say that the data set is normal.

It is OK for the data's histogram bars to lie slightly outside of the normal curve. This is where the p-value comparison to α becomes useful.

2.6 Sampling Techniques

There are a large variety of sampling techniques available to public health researchers. These can be broadly categorised into non-random and random sampling methods.

In order to ensure that samples chosen are representative, the selection process is very important.

How participants are selected in a study may influence the validity, reliability and overall results of the study.

While randomisation is seen as the best way to sample participants, it is not always practical or cost effective. Some bias is permissible depending on the type of study that is being done.

2.6.1 Sampling Bias

Sampling bias occurs when the sample that has been chosen is not representative of the larger population. Sampling bias can be classified into two broader terms information bias and selection bias.

Information bias occurs when there are errors in the measurements of collecting data.

- Observer bias – investigator's pre-existing knowledge may influence the way information is collected for the study.
- Interviewer bias – interviewer's may ask leading questions which steers the interviewees response to a certain direction.
- Response bias – participants may not answer questions truthfully when provided a survey.
- Recall bias – participants may not remember key information when asked questions.
- Instrument bias – equipment used for measuring is not calibrated properly.

2.6.1 Sampling Bias

Sampling bias can lead to skewed results or the study being unreliable or invalid.

Selection bias occurs when the researcher decides who is going to be studied. This often uses a non-random sampling method.

- Sampling bias – when participants in a target population are more likely to be chosen over other participants in the same population.
- Allocation bias – when there is a systematic difference between groups in a controlled trial.
- Attrition bias – participants may pull out of a study, which may affect results.

2.6.2 Non-random Sampling

Non-random examples of sampling

- Volunteers – people who volunteer (ie. Self-enrol themselves into a study) may cause bias if their characteristics differ from other participants. Volunteers may have motivation or be paid for their time.
- Snowballing – one participant involved in the study identifies other potential participants to join the study. Can occur with studies of rare diseases or where it is difficult to find participants.
- Convenience sampling – readily-accessible groups such as health service clients, university students or indigenous family/community (for an indigenous population study).
- Judgement sampling – participants are selected based only on the researcher's knowledge and opinion.
- Quota sampling – population is divided in strata (or groups) and final sample is a certain proportion from each group.
- Referred cases – participants may be referred by their health professional.

2.6.2 Non-random Sampling

Advantages of non-random sampling

- Easy to administer.
- Time and cost effective compared to random sampling.
- Useful for studies with small sample sizes.

Disadvantages of non-random sampling

- Sample selected may not be representative of the population.
- Results skewed. Descriptive statistics may be valid with inside group but not able to infer anything about larger population.
- Could lead to researcher bias – researchers influence results through selection to get a certain outcome.

2.6.3 Random Sampling

Random sampling is considered the gold standard of sampling. This is because using randomisation ensures each participant selected has an equal opportunity of being selected. In public health practice, random sampling is not always feasible.

Some random sampling methods include:

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Multi-stage sampling
- Cluster sampling

2.6.3 Random Sampling

Simple random sampling

- Every individual has equal likelihood of being selected.
- No volunteering or chosen cooperation
- Need to identify sampling frame e.g. Australian electoral roll (due to compulsory voting) for adult study or disease registers from hospitals.
- Each individual is assigned a number.
- Numbers are selected randomly (usually through computational methods such as Excel)

2.6.3 Random Sampling

Systematic sampling

- Systematic yet still random.
- From a queue or list, nth person is selected.
- Queue cannot be arranged in some certain way but can be based on time of attendance at a health service.
- Hence, systematic sampling can take place without an initial sampling frame or list.
- Ordering of people must be random
- Starting point must also be randomly chosen eg. Every 5th person, first selection must be randomly chosen.

2.6.3 Random Sampling

Stratified sampling

- Used when disease is unevenly distributed through a population eg. Sample is divided into “strata” such as age groups.
- Can aim for different proportions between strata eg. Prostate cancer in men is more prevalent in older males than younger males.
- Require sufficient numbers with the condition due to less frequency in the groups.
- Still randomly selected into the strata.

2.6.3 Random Sampling

Multistage sampling

- Practically speaking, sample subjects may need to be obtained in a series of stages.
- Eg. You want to find out which hospitals are preferred by children. List of all Australian children would be difficult to obtain so you start with sampling of Australian states. Next stage might be to sample from various hospitals in those states and next stage may be children of these hospitals.
- Also important to ensure children selected are randomly chosen.
- Can also be used with strata such as metro/non-metro groups within the sample.

2.6.3 Random Sampling

Cluster sampling

- Sampling a whole population can be difficult as they can be quite dispersed such as in remote Australia.
- This increases complexity and cost of simple random sampling.
- Easier to cluster groups of people eg. a single household contains several people in one place.
- Cluster sampling is convenient but can add more problems. Eg. Clusters of people in a small area using health services—possible that they all have the same GP, so if study is about quality of care then this creates a bias and the full range of possible care quality is not seen.
- Cluster sampling also useful for disease distributed in clusters eg. Malaria in villages with/without stagnant water nearby.

2.7 Type I and Type II Errors

Type I and II errors

If we say that we are “95% confident” in our findings, well there is a chance that we will be wrong 5% (or 1 out of 20) of the time.

Therefore when drawing conclusions, we introduce a possibility of an error.

A type I error is when we reject the null hypothesis when it is true.

A type II error is when we accept the null hypothesis when it is false.

Type I Error	Type II Error
False Positive	**False Negative**
reject null hypotheses when it is true	Accept (fail to reject) null hypothesis when it is false

2.7 Type I and Type II Errors

Example:

Patient has been diagnosed with a condition

(H_0 is either true or false)

	Patient does not have the condition (H_0 is true)	**Patient does have the condition (H_0 is false)**
Tests show negative result (Fail to reject H_0)		
Tests show positive result (Reject H_0)		

Test results (Accept or Reject H_0?)

A blood sample is taken from the patient…

2.7 Type I and Type II Errors

Patient has been diagnosed with a condition

(H_0 is either true or false)

Test results (Accept or Reject H_0?)

	Patient does not have the condition (H_0 is true)	**Patient does have the condition (H_0 is false)**
Tests show negative result (Fail to reject H_0)	**No error**	
Tests show positive result (Reject H_0)		

If test results come back NEGATIVE and patient DOESN'T have the condition then…

2.7 Type I and Type II Errors

Patient has been diagnosed with a condition

(H_0 is either true or false)

	Patient does not have the condition (H_0 is true)	**Patient does have the condition (H_0 is false)**
Tests show negative result (Fail to reject H_0)	**No error**	
Tests show positive result (Reject H_0)	**Type I error**	

Test results (Accept or Reject H_0?)

If test results come back POSITIVE and patient DOESN'T have the condition then…

2.7 Type I and Type II Errors

Patient has been diagnosed with a condition

(H_0 is either true or false)

Test results (Accept or Reject H_0?)

	Patient does not have the condition (H_0 is true)	Patient does have the condition (H_0 is false)
Tests show negative result (Fail to reject H_0)	**No error**	
Tests show positive result (Reject H_0)	**Type I error**	**No error**

If test results come back POSITIVE and patient DOES have the condition then…

2.7 Type I and Type II Errors

Patient has been diagnosed with a condition

(H_0 is either true or false)

Test results (Accept or Reject H_0?)

	Patient does not have the condition (H_0 is true)	Patient does have the condition (H_0 is false)
Tests show negative result (Fail to reject H_0)	**No error**	**Type II error**
Tests show positive result (Reject H_0)	**Type I error**	**No error**

If test results come back NEGATIVE and patient DOES have the condition then…