AUSTRALIAN CATHOLIC UNIVERSITY

PUBH620 Biostatistics

Life in Statistics

Week 1

Introduction to Biostatistics and Revision of Basic Concepts

ELECTRONIC WARNING NOTICE

Commonwealth of Australia
Copyright Act 1968
Form of notice for paragraph 135KA (a) of the *Copyright Act 1968*
Warning
This material has been copied and communicated to you by or on behalf of *Australian Catholic University under Part VA of the Copyright Act 1968 (the* ***Act).***
The material in this communication may be subject to copyright under the Act. Any further copying or communication of this material by you may be the subject of copyright or performers' protection under the Act.
Do not remove this notice.

Meet your lecturer

Dr Brandon Cheong

Senior Lecturer in Digital Health

Health, Science and Engineering

Research interests:

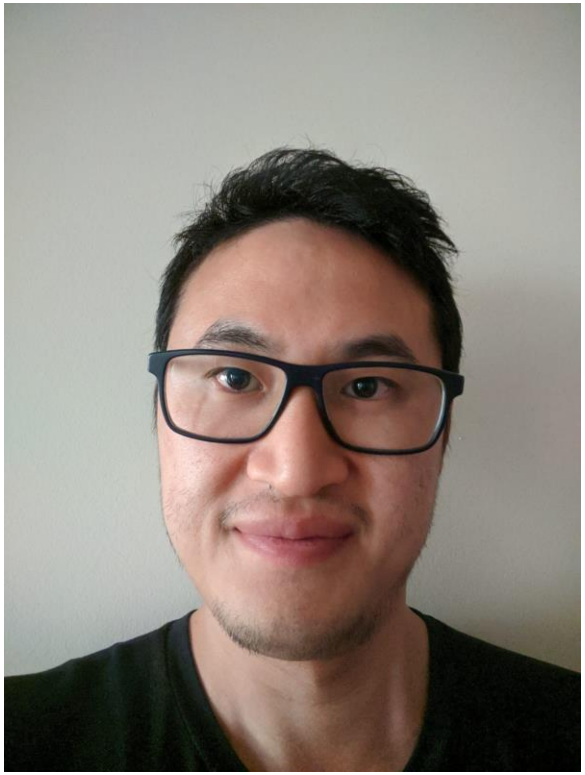
Health data standards

Health Interoperability

Consumer/clinician experiences

Artificial Intelligence in Health

Wearables technology



Meet your lecturer

Dr Govinda Poudel

Senior Lecturer

Health, Science and Engineering

Research interests:

Health Data Analytics

Neural Engineering

Neuroimaging

Computational Neurosciences

Neuroreporting Software Systems

In this workshop…

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

1.1 Overview of PUBH620 Biostatistics

1.2 Descriptive Statistics

- 1.2.1 Central Measures of Tendency
- 1.2.2 Central Measures of Spread
- 1.2.3 Visual Distributions

1.3 Scales of Measurement

1.4 Types of Variables

1.5 Hypothesis Testing

Topic Learning Objectives (TLOs)

- Understand the expectations of PUBH620 and assessment tasks
- Understand the need for statistics in health science research
- Revise basic statistical concepts in a public health context
- Understand the steps for hypothesis testing and form an appropriate null hypothesis

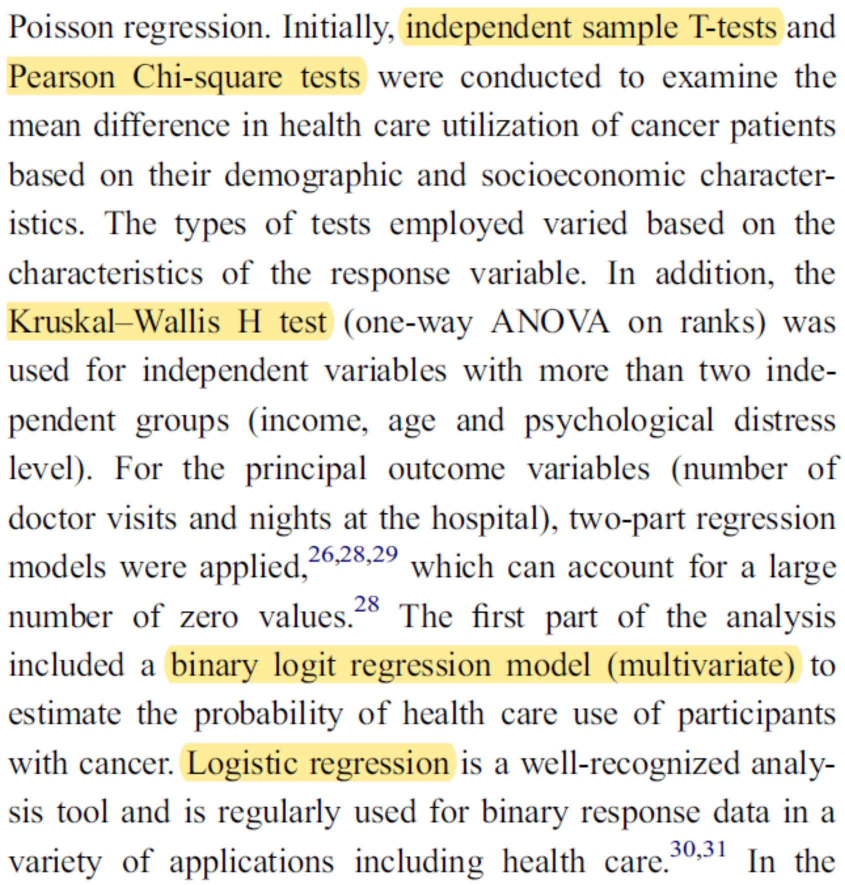
Public Health Applications of Biostatistics

Some applications include:

- Cancer treatments
- Disease control
- Surgery
- Drug testing
- Life science applications

Eg. Research Article

Rana, R. H., Alam, K., Gow, J., & Ralph, N. (2019). Predictors of health care use in Australian cancer patients. *Cancer management and research*, *11*, 6941.



Poisson regression. Initially, independent sample T-tests and Pearson Chi-square tests were conducted to examine the mean difference in health care utilization of cancer patients based on their demographic and socioeconomic characteristics. The types of tests employed varied based on the characteristics of the response variable. In addition, the Kruskal–Wallis H test (one-way ANOVA on ranks) was used for independent variables with more than two independent groups (income, age and psychological distress level). For the principal outcome variables (number of doctor visits and nights at the hospital), two-part regression models were applied,[26,28,29] which can account for a large number of zero values.[28] The first part of the analysis included a binary logit regression model (multivariate) to estimate the probability of health care use of participants with cancer. Logistic regression is a well-recognized analysis tool and is regularly used for binary response data in a variety of applications including health care.[30,31] In the

(Rana et al., 2019)

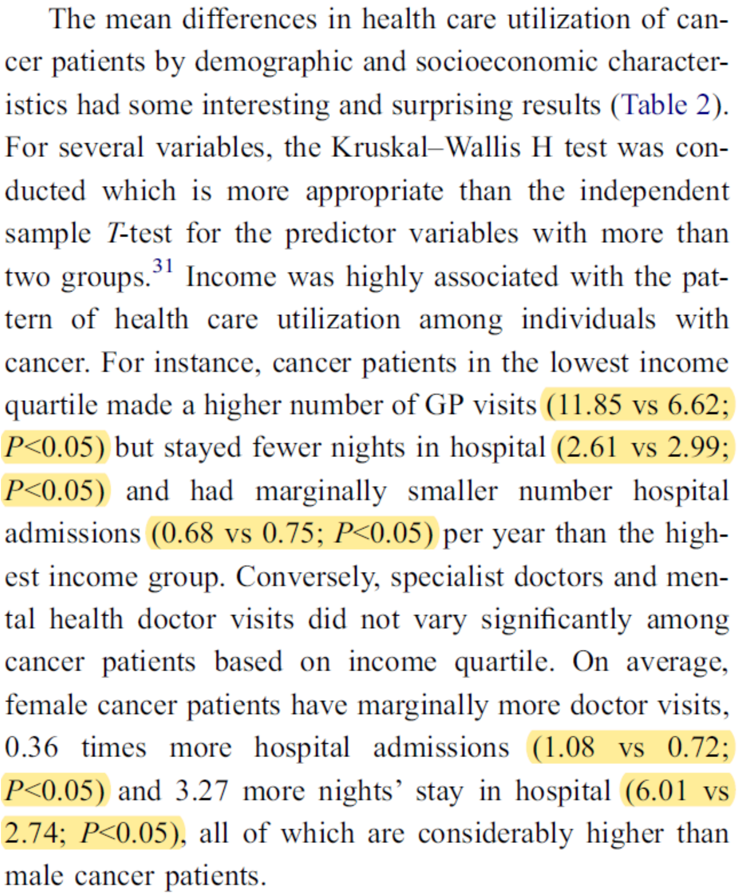
Public Health Applications of Biostatistics

Statistics in health journal articles

Rana, R. H., Alam, K., Gow, J., & Ralph, N. (2019). Predictors of health care use in Australian cancer patients. *Cancer management and research*, *11*, 6941.

What does it all mean?

You will learn this over the next 12 weeks!



The mean differences in health care utilization of cancer patients by demographic and socioeconomic characteristics had some interesting and surprising results (Table 2). For several variables, the Kruskal–Wallis H test was conducted which is more appropriate than the independent sample *T*-test for the predictor variables with more than two groups.[31] Income was highly associated with the pattern of health care utilization among individuals with cancer. For instance, cancer patients in the lowest income quartile made a higher number of GP visits (11.85 vs 6.62; *P*<0.05) but stayed fewer nights in hospital (2.61 vs 2.99; *P*<0.05) and had marginally smaller number hospital admissions (0.68 vs 0.75; *P*<0.05) per year than the highest income group. Conversely, specialist doctors and mental health doctor visits did not vary significantly among cancer patients based on income quartile. On average, female cancer patients have marginally more doctor visits, 0.36 times more hospital admissions (1.08 vs 0.72; *P*<0.05) and 3.27 more nights' stay in hospital (6.01 vs 2.74; *P*<0.05), all of which are considerably higher than male cancer patients.

(Rana et al., 2019)

1.1 Overview of PUBH620 Biostatistics

This unit is very difficult! If you do not keep up with the weekly material, you will surely fail.

Each concept builds on what came before.

If you miss a week of content, make sure you catch up as soon as possible!

Essential skills that employers look for.

You will learn to use and understand statistics and its relevancy in public health.

Software

You must have access to SPSS and Microsoft Excel to complete the exercises of this unit.

Please make sure you have working copies of each on your computer. SPSS can be purchased for $45 and ACU provides a free copy of Microsoft Office to students.

Always remember to back up your work!!! Use a USB drive, cloud storage system (ACU OneDrive), Google Drive, Drop Box etc.

Week 1 – Introduction to Excel

Week 2 – Introduction to SPSS

Assessment Tasks

Assessment Task 1: Analysis of a simulated data set

Statistical methods written assignment – 20%

Assessment Task 2: Preparation of statistical methods and analysis for peer-reviewed journal article

Write up of Methods, Results, Discussion and Conclusion for public health scenario – 40%

Assessment Task 3: Assuming the role of a statistician

Conversation with your lecturer about statistical methods for public health research – 40%

1.2 Descriptive Statistics

Descriptive statistics are tools that allow us to summarise a data set.

For a data set, we are often concerned with its central measure, variability and distribution.

Central measures of tendency

- Mean
- Median
- Mode

Central measure of spread (variability)

- Range
- Standard Deviation
- Variance

Visual Distributions

- Frequency tables
- Boxplot
- Histogram
- Bar chart
- Scatterplot

1.2.1 Central measures of tendency

Population mean

Most common measure of centre. The **population mean** denoted by μ:

$$\mu = \frac{\sum x}{N} = \frac{\text{sum of the scores}}{\text{population size}}$$

Eg.

Data set	**10**	**23**	**56**	**9**	**77**	**38**	**101**

Mean = 44.86

Excel tip: use “=average” not “=mean” to calculate the mean of a data set.

1.2.1 Central measures of tendency

Sample mean

Most common measure of centre. The **sample mean** denoted by $\boldsymbol{\bar{x}}$ (*x* bar):

$$\bar{x} = \frac{\sum x}{n} = \frac{\text{sum of the scores}}{\text{sample size}}$$

Eg.

Data set

10	**23**	**56**	**9**	**77**	**38**	**101**

Mean = 44.86

Still calculated in the same way but for a sample!

The mean is greatly affected by outliers! Ie. Add 2735 to the data set and the mean becomes 381.125, which is a lot higher than the original mean we calculated.

1.2.1 Central measures of tendency

Median

The middle number when all data are in order:

Eg. What is the median of the below data set?

12 11 10 4 3 8 5 4 11 5 7 4 6 7 9

Reorder number from lowest to highest (or highest to lowest)

3 4 4 4 5 5 6 7 7 8 9 10 11 11 12

The middle number is 7…note that we have 15 numbers, which is an odd number. What if we had even numbers?

1.2.1 Central measures of tendency

Median

The middle number when all data are in order:

Let’s make it 16 numbers, an even number

12 11 10 4 3 8 5 4 11 5 7 4 6 7 9 3

We still reorder numbers from lowest to highest (or highest to lowest)

3 3 4 4 4 5 5 6 7 7 8 9 10 11 11 12

The middle numbers are 6 and 7, the median is the average of these two numbers ie. (6+7)/2 = 6.5

Excel tip: use “=median” to calculate median in Excel.

1.2.1 Central measures of tendency

Mode

The most common number that appears in the data set:

Person	Rey	Kylo	Finn	Poe	Chewbacca
No. porgs eaten	3	2	3	1	3

Mode = 3

Excel tip: use “=mode” to calculate mode in Excel

1.2.1 Central measures of tendency

When it comes to descriptive statistics, the central measures of tendency are not enough…

Eg. Consider the four data sets below…calculate the means of each data set…what do you get?

	Data set 1	Data set 2	Data set 3	Data set 4
	20	15	5	1
	20	28	7	2
	20	19	14	4
	21	23	33	5
	21	17	43	90
mean				

As we move from data set 1 to data set 4 we see that spread increases. Hence we need a measure of spread…central measures of spread!

1.2.2 Central measures of spread

Range

The difference between the largest and smallest values in a data set:

3 4 4 4 5 5 6 7 7 8 9 10 11 175

Range = 175 - 3 = 172

Just like the mean, the range is also greatly affected by outliers.

3 4 4 4 5 5 6 7 7 8 9 10 11 175 1200

Range = 1200 - 3 = 1197

Excel tip: use “=range” to calculate the range in Excel.

1.2.2 Central measures of spread

Standard Deviation

The most commonly used measure of spread. Measures the average amount by which numbers in data set differ from the mean.

Standard deviation (SD) is calculated differently for a population and sample.

σ denotes population SD – use “=stdev.p” in Excel

s denotes sample SD – use “=stdev.s” in Excel

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \qquad s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Red circle part is known as the Sum of Squares. A key statistical concept that needs to be learnt.

Notice a difference between the two formulas?

1.2.2 Central measures of spread

Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \qquad s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Why is n- 1 used here and not n? This is due to Bessel's correction (aka degrees of freedom), we understand that the sample size will always be less than the population size so in order to give a more accurate representation of the SD, we use what's called a degrees of freedom, which is n – 1. This reduces bias in our calculation.

Note: the sum of squares divided by the degrees of freedom is known as the variance.

1.2.2 Central measures of spread

Standard Deviation

Think of degrees of freedom like this…
You have 7 different chocolate bars, that's one chocolate bar to be eaten a day (7 day week)!
On the first day, you can choose any of the 7 chocolate bars you have to eat.
On the second day, you only have 6 remaining chocolate bars to choose from…
On day 3 you only have 5 chocolate bars left to choose from…
On day 4, you only have 4 chocolate bars to choose from
Day 5, 3 bars left, Day 6, 2 bars left….and Day 7, no choice but the 1 bar left.

Degrees of freedom in statistics is simply the number of observations that can vary when determining statistical parameters. You had 7-1 = 6 days of choice of chocolate bars. Hence n – 1.

Note: this is not an overly important concept to know but you will see it throughout the semester!

1.2.2 Central measures of spread

Variance

Square the standard deviation and you get the variance. A very important measure of spread. Hence:

σ^2 denotes population variance – use “=var.p” in Excel

s^2 denotes sample variance – use “=var.s” in Excel

More often than not you will work with samples instead of populations.

Sample Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Note: Variance is used frequently in hypothesis testing! Make sure you know it!

1.2.3 Visual distributions

Boxplot

Boxplots are good for visualising descriptive statistics.

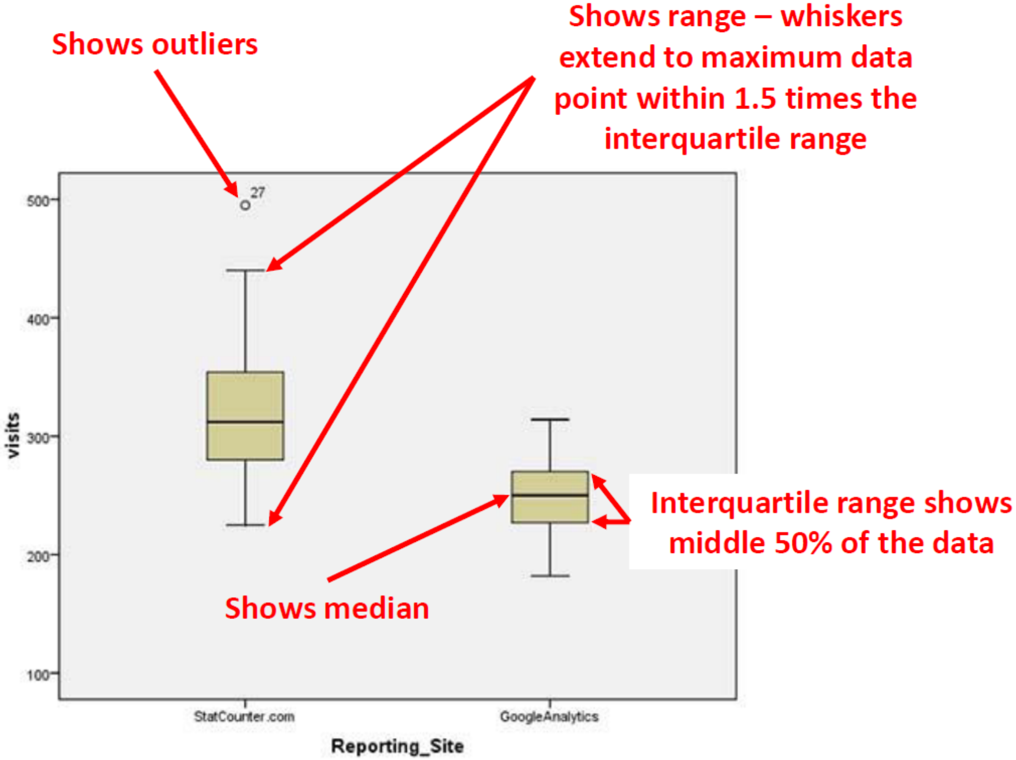
Shows:

Median

Range

Outliers

Comparisons



Shows outliers

Shows range – whiskers extend to maximum data point within 1.5 times the interquartile range

Interquartile range shows middle 50% of the data

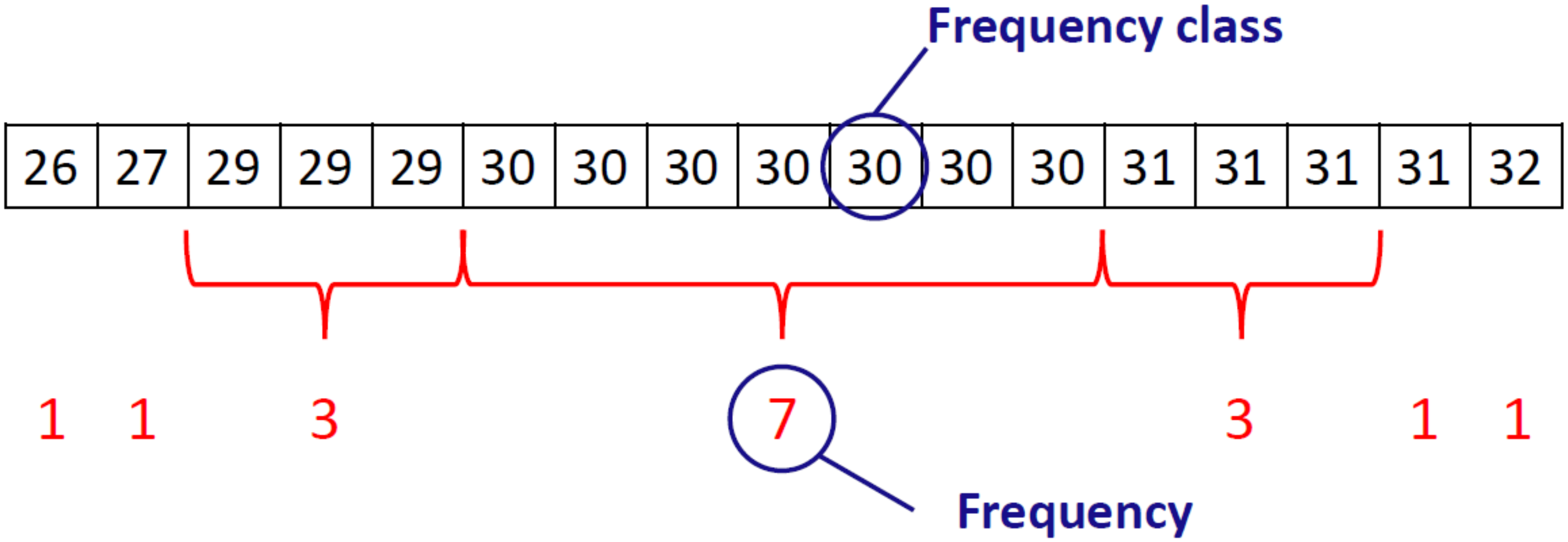
Shows median

1.2.3 Visual distributions

Histogram

Histograms are good for visualising frequency data and distributions.

Eg. Foot length of male PUBH620 students (in cm).



Frequency class

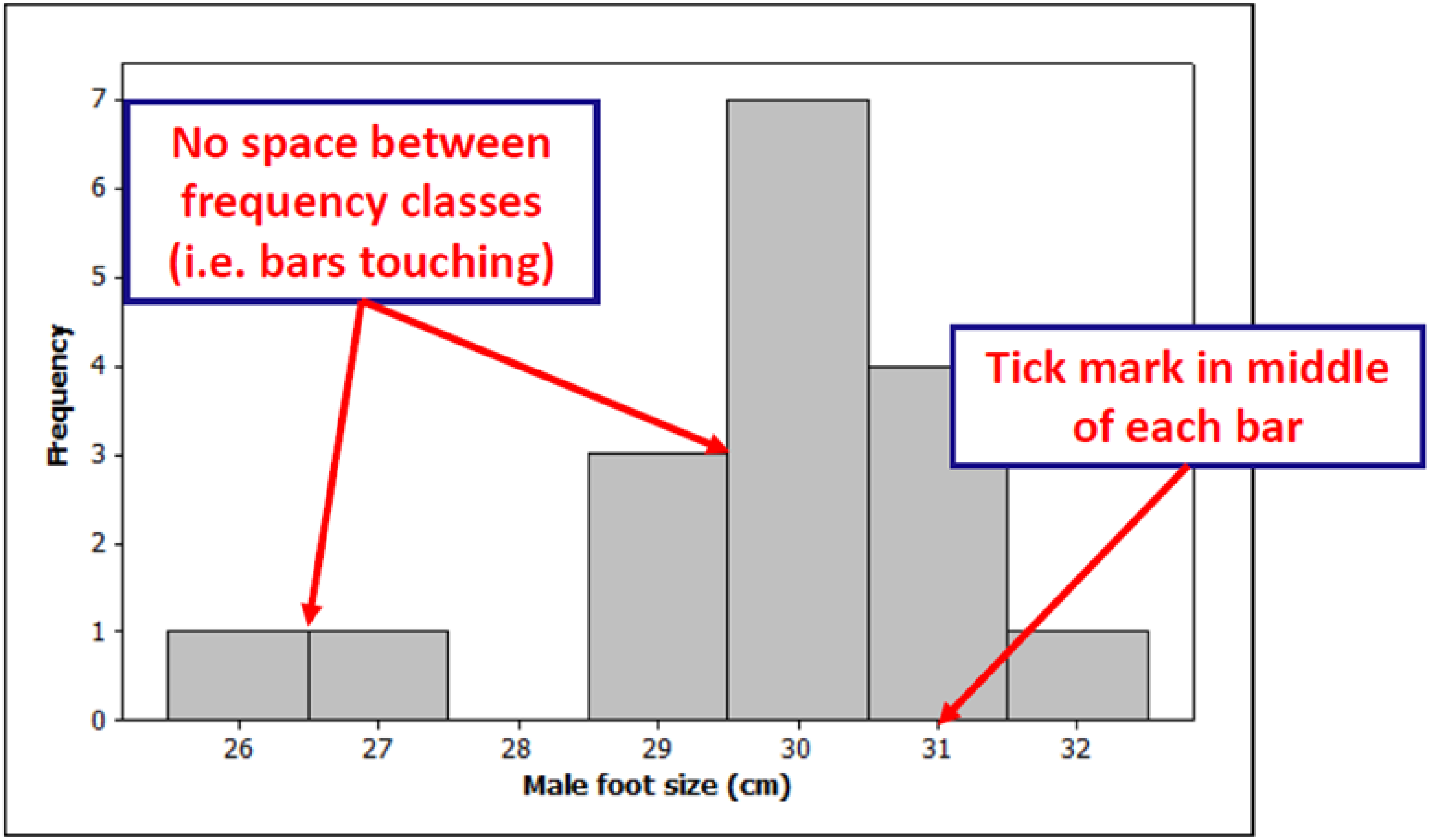
26	27	29	29	29	30	30	30	30	30	30	30	31	31	31	31	32

1 1 3 7 3 1 1

Frequency

1.2.3 Visual distributions

Histogram



No space between frequency classes (i.e. bars touching)

Tick mark in middle of each bar

1.2.3 Visual distributions

Histogram

Histograms will ideally have 5 – 10 frequency classes. If there are too many frequency classes, divide them in class intervals.

Class intervals must be:

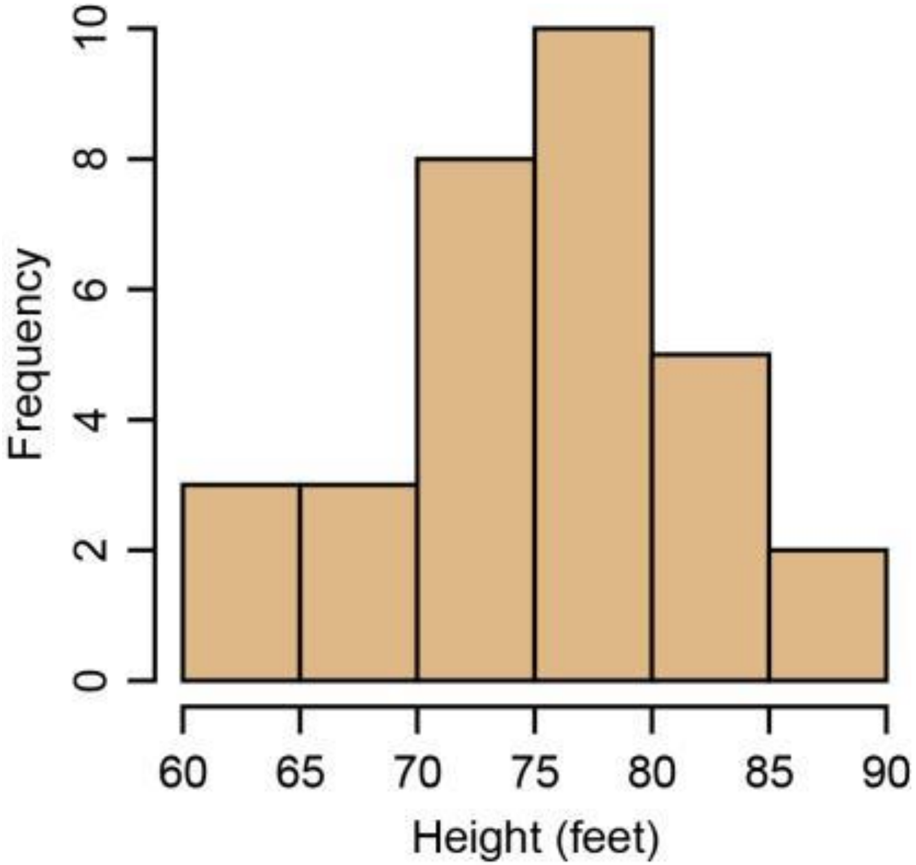
- **Same size**
- **Exclusive – datum only occurs in one interval**

Eg.

Height (feet)	60 - <65	65 - <70	70 - <75	75 - <80	80 - <85	85 - <90
Frequency	3	3	8	10	5	2

1.2.3 Visual distributions

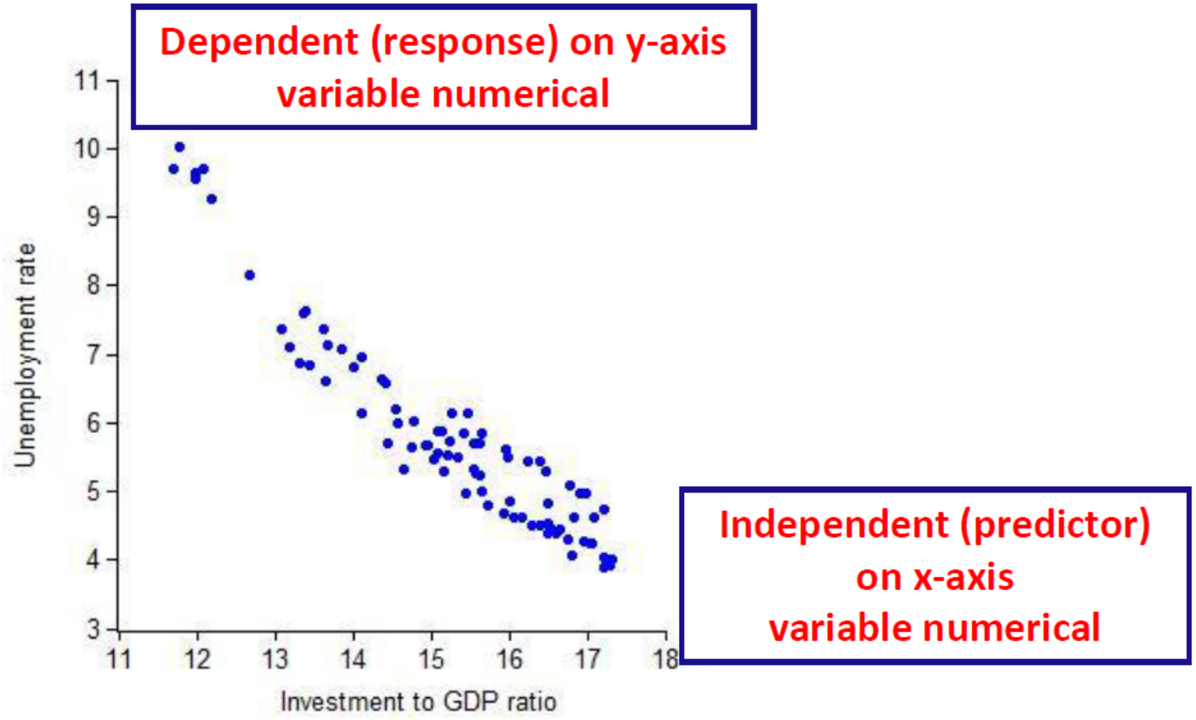
Histogram



1.2.3 Visual distributions

Scatterplot

Scatterplots are good for looking at the relationships between two variables.



Dependent (response) on y-axis
variable numerical

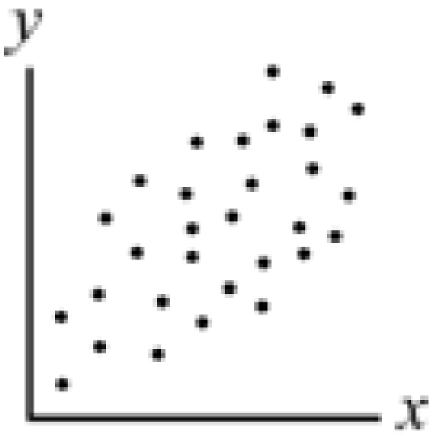
Independent (predictor)
on x-axis
variable numerical

1.2.3 Visual distributions

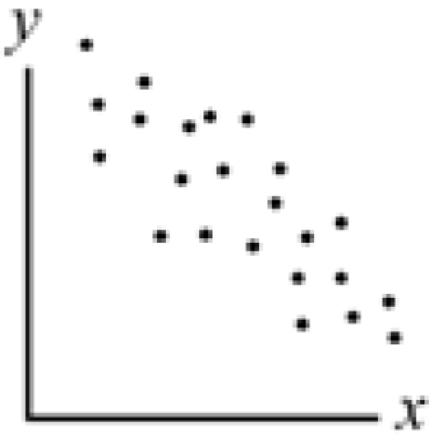
Scatterplot

Relationship between the two variables is usually linear and either positive or negative.

Useful for exploring relationship between two variables



positive



negative

1.2.3 Visual distributions

Bar chart

Bar charts contain one categorical variable usually on the x-axis.

Eg.

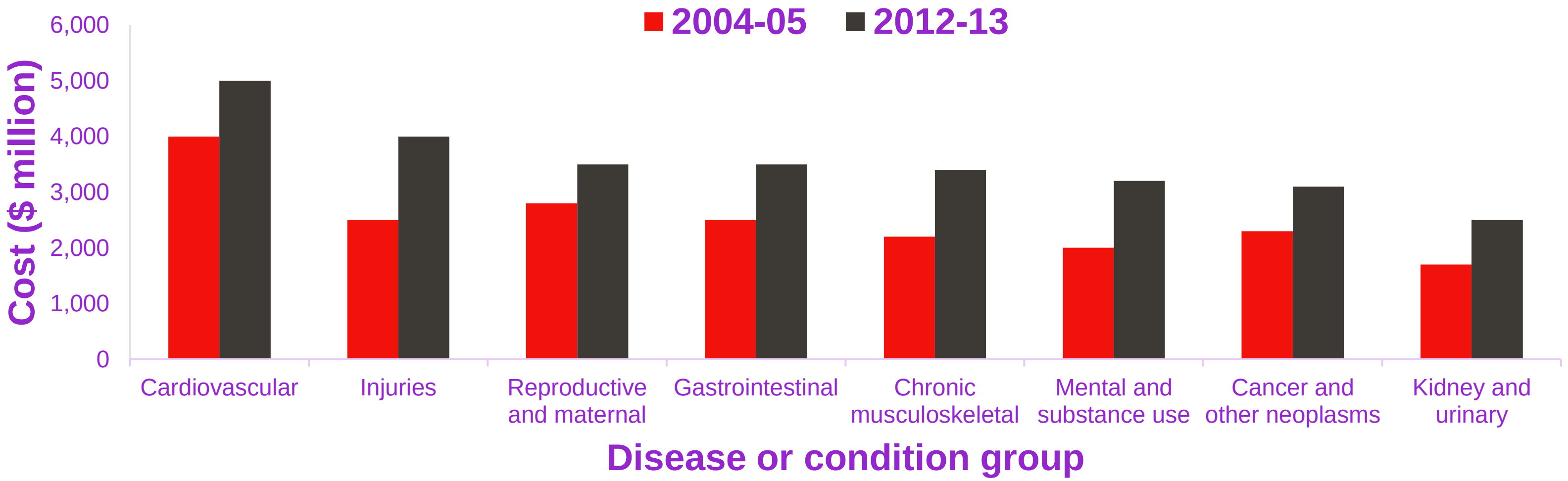


Figure 1. Australian Health Expenditure on Disease or Condition Group in 2004-05 and 2012-13. Data obtained from the Australian Institute of Health and Welfare “Australia’s health 2016 – in brief”.

1.2.3 Visual distributions

Good graphs

It is very important to present your graphs correctly!

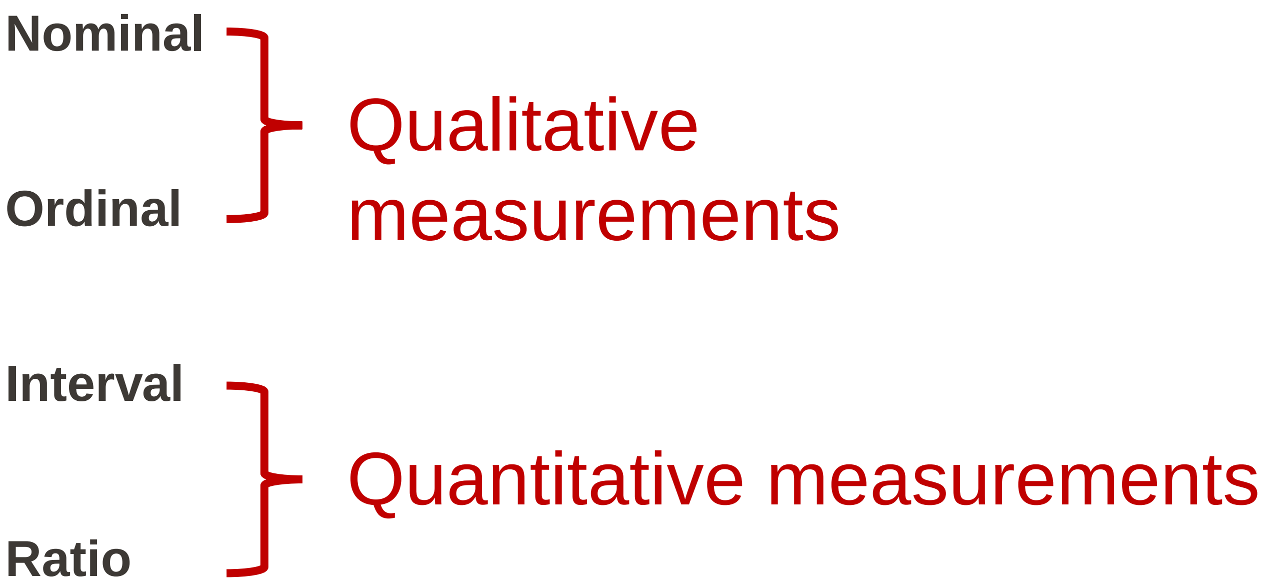
Good graphs…

- Are simple, not cluttered, and readable
- Have the independent variable on the x-axis
- Have labels on the axes (with units if needed)
- Have a legend if there is > 1 treatment group
- Have appropriate number scales
- Have a caption (below) that explains the figure

1.3 Scales of measurement

Scales of measurement

A very important concept to know. Knowing the scale of measurement helps to identify the type of statistical test you will choose. Note: SPSS treats Interval and Ratio as the same.



Nominal

Ordinal

Qualitative measurements

Interval

Ratio

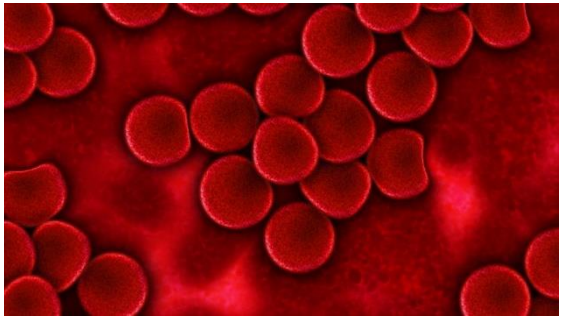
Quantitative measurements

1.3 Scales of measurement

Nominal scale

Classifies objects or events in to categories or names.

Eg.



Marital status: single, married, de facto etc.

Sex: male and female

Blood group: A, B, AB, O

Type of disease: Type I or Type II diabetes, Stages of cancer (0 – IV)

- Ranking or order is not important
- Can be dichotomous/binary (male or female, smoker or non-smoker)
- Mutually exclusive

1.3 Scales of measurement

Ordinal scale

Classifies and **ranks** experimental units measured.

Eg. Level of pain a patient experiences.

None

Mild

Moderate

Severe

Very severe

Worst possible

Ranked by increasing pain

- Does not indicate absolute values
- Does not assume equal intervals between levels
- Does not describe relationships between individuals in different classifications

1.3 Scales of measurement

Ordinal scale

Eg. A survey asks the patient their pain levels using a Likert-scale. (On a scale of 0 to 5, how much pain are you experiencing today?) Allocate scores to each pain level (Likert-scale)

None = 0

Mild = 1

Moderate = 2

Severe = 3

Very severe = 4

Worst possible = 5

Mathematical operations cannot be performed for ordinal numbers. For eg. The difference between scores 4 and 5 would not be equal in pain magnitude to scores of 0 and 1. Clinically speaking, the intensity of pain for score 2 (moderate) would not be twice that of score 1 (mild).

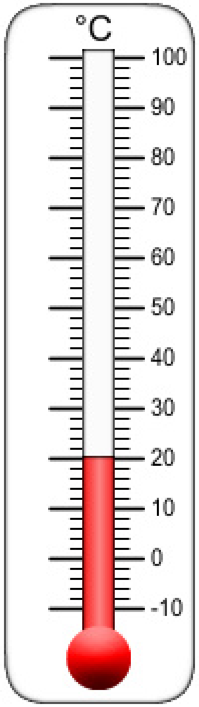
1.3 Scales of measurement

Interval scale

Classifies and **ranks** and has **fixed interval**.

Eg. Temperature in °C.

- Fixed interval; can add and subtract
  - 10° and 20° differ by 10°
  - 40° and 60° differ by 20°
- No absolute zero, cannot divide
  - 0°C does not = no heat energy present
  - 30°C is not 3 times hotter than 10°C



Note: Temperature is ratio scale when measured in Kelvin…why?

1.3 Scales of measurement

Ratio scale

Classifies and **ranks,** can have **fixed intervals** and an **absolute zero**.

Eg. Measures of weight, height, time etc...

- Fixed interval; can add and subtract
  - 2 minutes + 2 minutes = 4 minutes
- Absolute zero; can multiply or divide
  - 0 minutes indicates no time lapsed
  - 2 minutes is twice as long as 1 minute

1.3 Scales of measurement

Summary of scales of measurement

Scale	Nominal	Ordinal	Interval	Ratio
Classifies	✓	✓	✓	✓
Ranks		✓	✓	✓
Interval			✓	✓
Absolute zero				✓

More information

1.3 Scales of measurement

Converting scales

Patient ID	A	B	C	D	E
Height (m)	1.87	1.72	2.09	1.67	2.13

Convert to ordinal scale

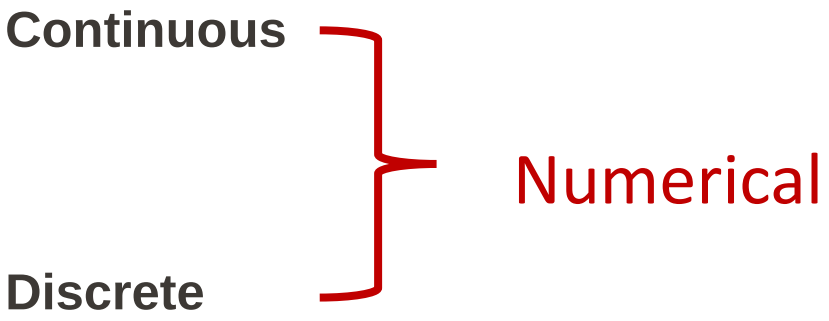
Patient ID	A	B	C	D	E
Rank	3	4	2	5	1

What information is lost?

Can still see which patient is shorter,
but not HOW MUCH shorter

1.4 Types of variables

Categorical Non-numerical



Continuous

Discrete

Numerical

1.4 Types of variables

Categorical Non-numerical

Continuous

Numerical

Discrete

Note: Do not confuse “type of variables” with “scales of measurement”. A common mistake!

1.4 Types of variables

Categorical variable

Categorical variables are fit in to distinct categories.

Eg.

Variable	Categories
Sex	Male, Female, Intersex
Marital status	Single, Married, De Facto
Blood type	A, B, AB, O
Blood glucose level	Low, Medium, High

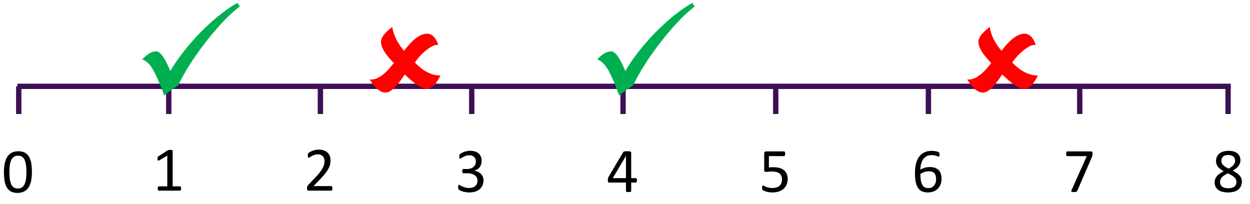
Often uses descriptive words to categorise the variable.

1.4 Types of variables

Discrete variable

Discrete variables can only take values of equal to whole numbers or integers.

Eg. Number of children in a family, number of visits to the hospital etc.



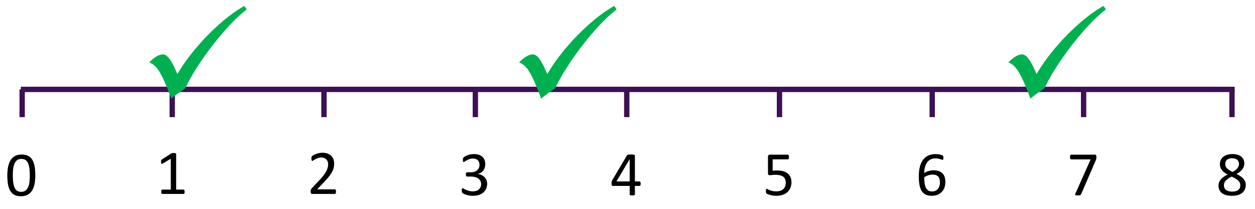
You cannot have 2.337 children admitted to hospital…this would be inhumane!

1.4 Types of variables

Continuous variable

Continuous variables are measurements that can take decimal values. Values can exist along a continuum.

Eg. Height, weight, BMI, blood pressure, blood glucose level, cortisol levels etc.



Keep in mind that continuous variable can also be expressed as whole numbers.

1.5 Hypothesis testing

Hypothesis testing

Hypothesis testing is a statistical method used to help us make statistical decisions about data we have obtained.

Based on an observation we might have for an event, a hypothesis is the best explanation possible from the facts we possess. It is also a statement that is testable (hence used for conducting experiments) and a very important concept in statistics.

General steps for doing a hypothesis test 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

1.5 Hypothesis testing

Hypothesis testing example

Cyanosis is a condition where a person’s blood is low in oxygen levels. As a result a bluish colour is present to their skin.

You observe a person from afar and they happen to have a blue tinge to their skin. Could they have cyanosis?

You hypothesise that the person DOES have cyanosis.

How would you test this? What are other causes of blue skin?

1.5 Hypothesis testing

Hypothesis testing example – other causes of blue colouration?

Are they a na’vi?

1.5 Hypothesis testing

Hypothesis testing example – other causes of blue colouration?

Are they cosplaying?

1.5 Hypothesis testing

Hypothesis testing example – other causes of blue colouration?

Are they a smurf?

1.5 Hypothesis testing

Hypothesis testing example

We aren’t certain that the person definitely has cyanosis so we need to…

Test our hypothesis!

1.5 Hypothesis testing

Hypothesis testing

Hypothesis testing generally states that there is a (statistical) difference between groups, or a relationship between variables, or that there has been a change in something.

With so many “changes” taking place, it is often impossible to prove our hypothesis and therefore we need to form what’s called a **null hypothesis.**

A null hypothesis is a testable statement that can be falsified. So instead of taking the form of difference, relationship and change…

We take the form of NO difference, NO relationship and NO change.

The Null Hypothesis is denoted by H_0

The actual hypothesis is the Alternate Hypothesis and is denoted by H_1

1.5 Hypothesis testing

Hypothesis testing

Let’s temporarily step away from our blue person example and focus on the null hypothesis using swans!

You are a british bird watcher and observe that all swans you have seen are white.

1.5 Hypothesis testing

Hypothesis testing

So you hypothesise that all swans everywhere in the world are white!

How do you prove your hypothesis?

The only way to do this is to observe all swans everywhere and see that they are white.

But this is impossible!

Furthermore…you actually see a black swan in Australia!

1.5 Hypothesis testing

Hypothesis testing

Our actual hypothesis (H_1) was that all swans were white everywhere in the world but as this is impossible to test, we state a null hypothesis (H_0) that “not all swans in the world are white”…this is easier to prove true than our actual hypothesis.

1.5 Hypothesis testing

Hypothesis testing

Since we cannot prove that the alternate hypothesis (**H_1**) is true

We try to show that the **H_0** is unlikely to be true

If the **H_0** does not match the facts, we say that there is evidence to support the **H_1**

A tricky concept that takes time to sink in

1.5 Hypothesis testing

Hypothesis testing

Back to our blue person…

You observe a person from afar with a blue colour to their skin

Alternate hypothesis (**H_1**): the person has cyanosis

Null hypothesis (**H_0**): the person does not have cyanosis

You test H_0 and their oxygen level is very low (below 94%)

Therefore you have evidence to support the H_1