Understanding the Data!
What do I need to know about statistics?

Prof Dominique Cadilhac

Contributions from Tara Purvis

Stroke and Ageing Research

Department of Medicine, School of Clinical Sciences at Monash Health

Head: Stroke Division, Florey Institute of Neuroscience and Mental Health

dominique.cadilhac@monash.edu

@DominiqueCad

Outline

- Overview of common statistics for publications with quantitative data

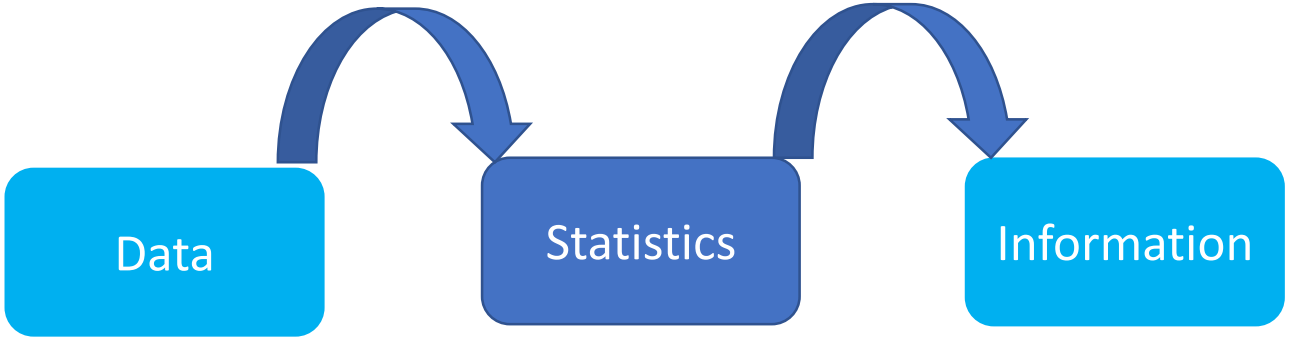
  (may be we will do qualitative analytic methods another time! ☺)

- Understanding and interpreting data
- Aspects of critical appraisal
- Examples from Australian Stroke Clinical Registry and National Audit

What are statistics? METHODS of ANALYSIS

The process to organise, analyse and interpret data

- Tool for converting data in to understandable quantitative information



Statistics and the research process

1. Research questions
2. Hypotheses

"If _____[I do this] _____, then _____[this]_____ will happen."

Testable prediction about what will happen

A good hypothesis defines the variables in easy-to-measure terms, e.g. the during the testing, and what the effect of the changes will be

3. Study design
4. Collecting data and checking quality
5. Analysing data (***statistics)***
6. Interpreting results

www.sciencebuddies.org/blog/a-strong-hypothesis

A **hypothesis is** defined as an educated guess, while a **research question is** simply the researcher wondering about the world

Do men receive thrombolysis more than women?

or

It is hypothesized that because more men use ambulance transport for stroke than women they receive thrombolysis more often.

https://sciencing.com/the-difference-between-research-questions-hypothesis-12749682.html

What is the null hypothesis?

Default position that there is nothing new happening, *(i.e. no association among groups, or no relationship between two measured phenomena)* and no statistical significance will be found

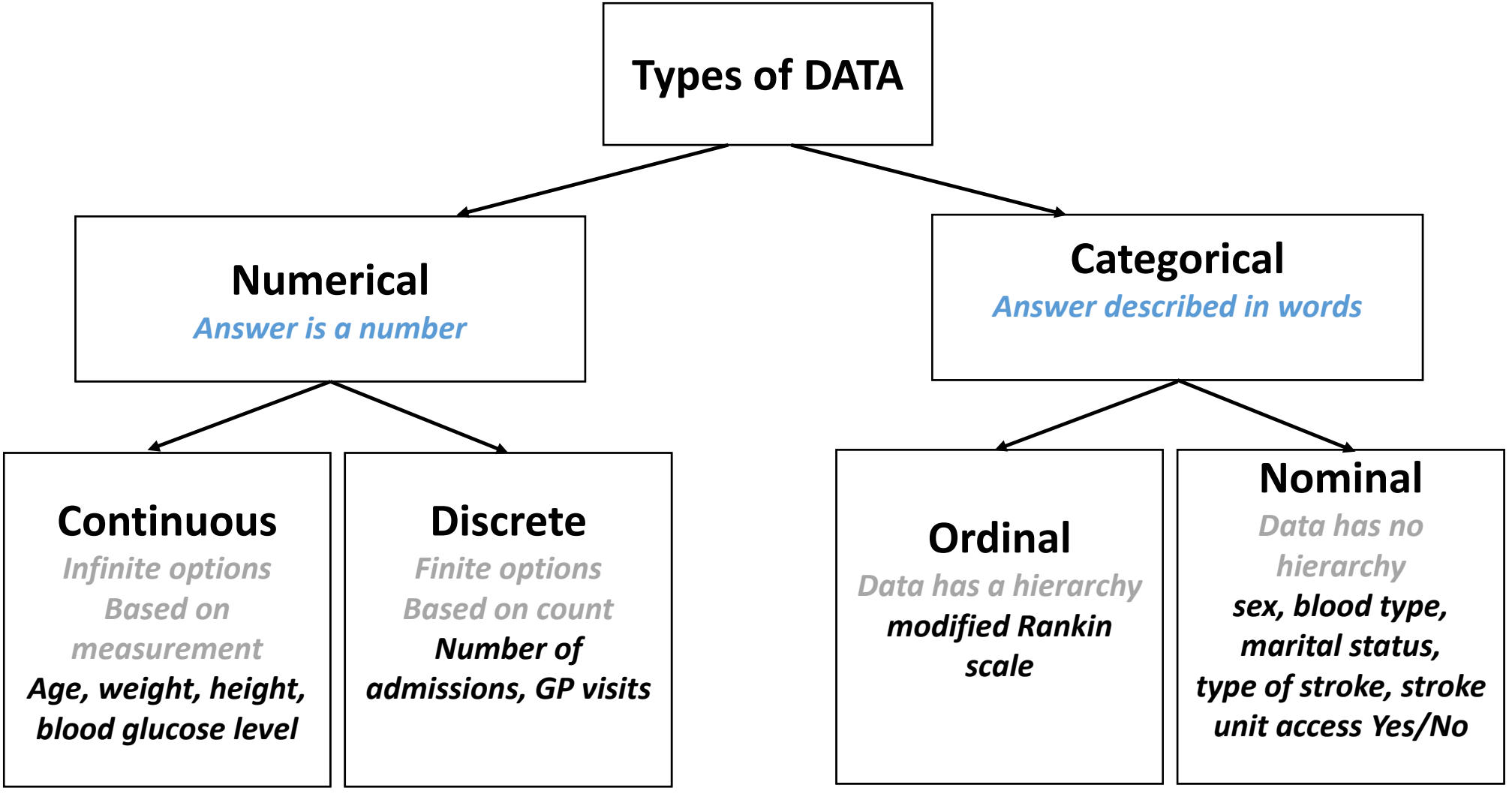
Any observed difference is due to sampling or experimental error

e.g. Men and women have an equal chance of receiving thrombolysis

Publications of original research

Methods:

- Study design: observational or randomised controlled comparison
- Data collection
  - Definitions
  - Single site, multisite, or community based
  - Outcome assessments: self-reported or interviewed and “blinded” if an RCT
- Data analysis:
  - Descriptive with univariable statistics
  - Multivariable (*where there is adjustment for several factors that influence outcome e.g. age, sex, stroke type*)
  - Sensitivity analyses: alternate inputs or factors to assess influence on the outcome of the analysis


Types of DATA
Numerical
Answer is a number
Categorical
Answer described in words
Continuous
Infinite options
Based on measurement
Age, weight, height, blood glucose level
Discrete
Finite options
Based on count
Number of admissions, GP visits
Ordinal
Data has a hierarchy
modified Rankin scale
Nominal
Data has no hierarchy
sex, blood type, marital status, type of stroke, stroke unit access Yes/No

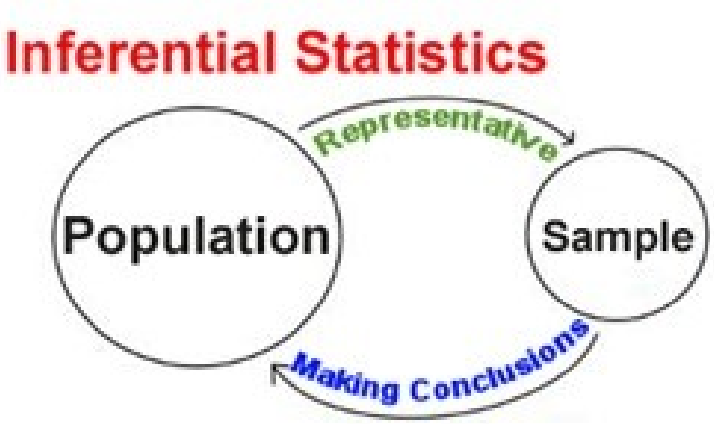
Statistics

Descriptive

collecting, organising, summarising, analysing and presenting data.

Inferential

Other important concepts:



Population: is any complete pool from which a statistical sample is drawn.

Sample: is a sub-set of the population you want to know something about.

Must be large enough to provide a reliable representation of the population and related ***random variation***

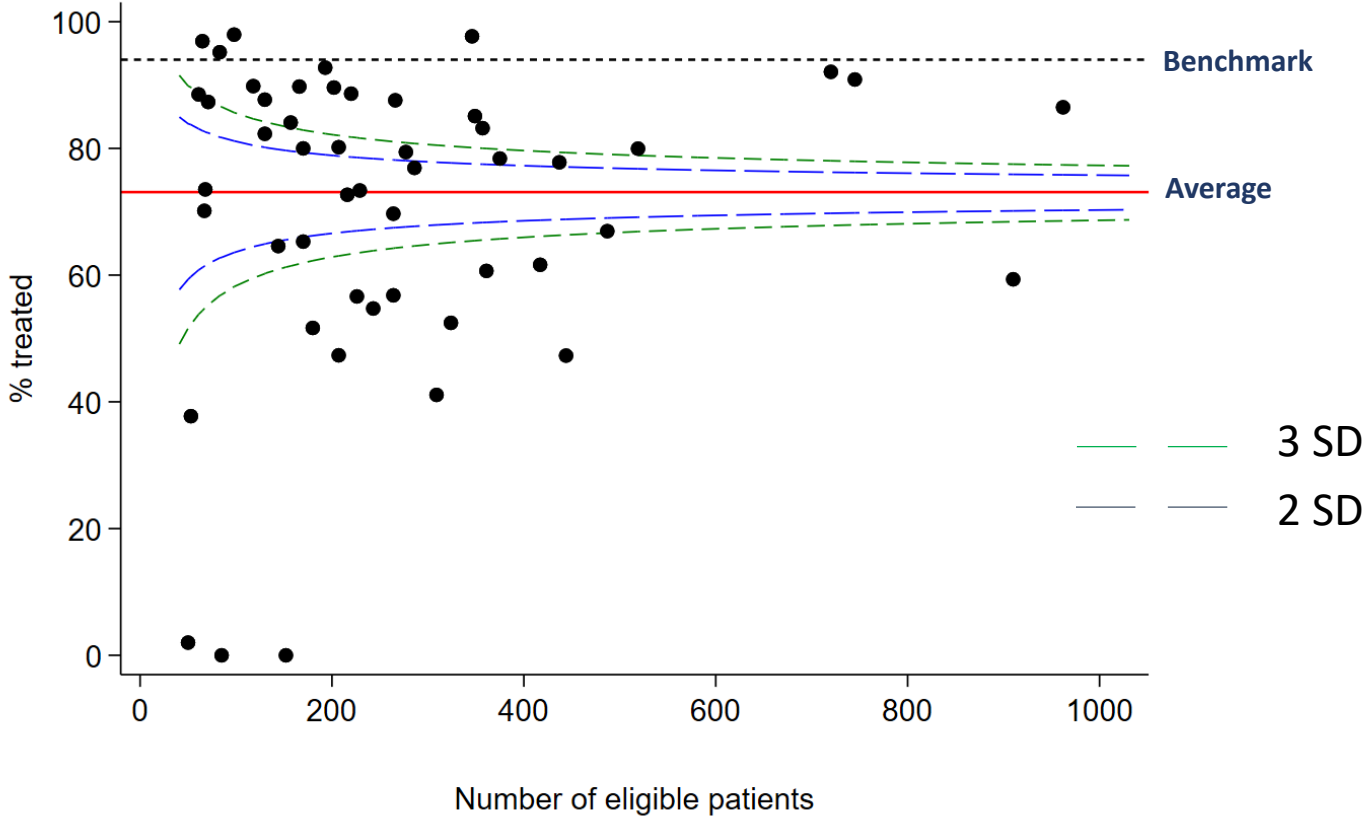
When we read a paper we need to understand if the sample was representative of the population: response rate, age groups, sex distribution

Example of samples of data

Received Stroke Unit Care

Overall mean adherence was 73% (red line)

Each dot represents adherence for a single hospitals (a cluster of data representing a source of 'truth')



SD: standard deviation

Data can be collected and presented different ways

modified Rankin Scale: mRS

Ordinal
0, 1, 2, 3, 4, 5, 6

Categorical
Independent = 0-2
Dependent = 3-5

Dichotomous

Takes on one of only two possible values

Age (years)

Continuous

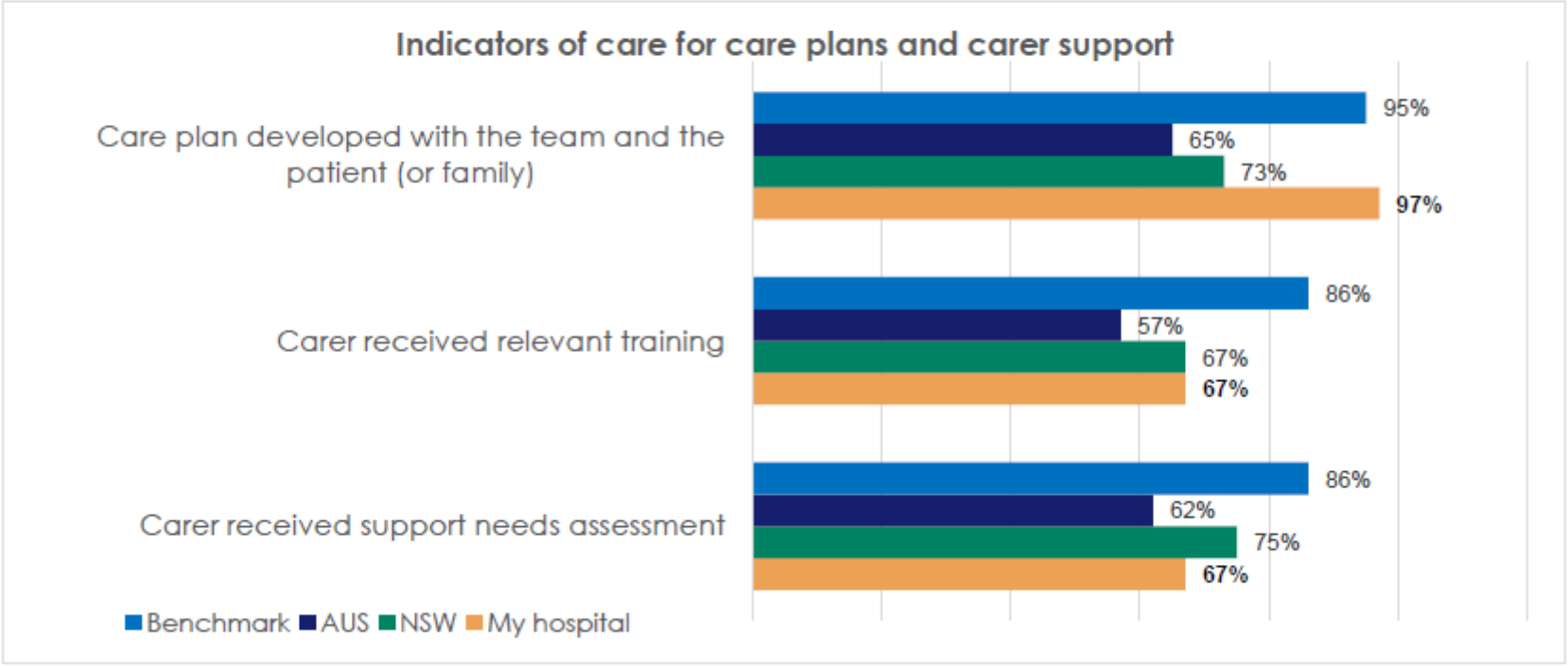
Under or over
65 years

Categorical
18-24
25-34
35-44
45-54
55-64
65+

Example data from the Stroke Foundation Audit Report

Care planning and carer support

Care plan developed with team and patient	State ranking 9 / 44	Changes over time 100% (2011) 73% (2013) 92% (2015)



Stroke FOUNDATION

2017 Audit Results

Descriptive statistics (univariable)

- Numerator/Denominator = %
- Inclusion and exclusion criteria very important
  - *Who is in and who is out?*
- Sample size: N

Example Mood assessment

Numerator: Patients who received a mood assessment
Denominator: Patients with stroke

If we assume not documented = NO

"Was mood assessed?"

Mood assessed	2018 Rehab Audit
Yes	2057 (56%)
No	765 (21%)
Not documented	829 (23%)

$$\frac{\text{Yes}}{\text{No + Yes}} = 73\% \quad \text{vs} \quad \frac{\text{Yes}}{\text{No + Yes + ND}} = 56\%$$

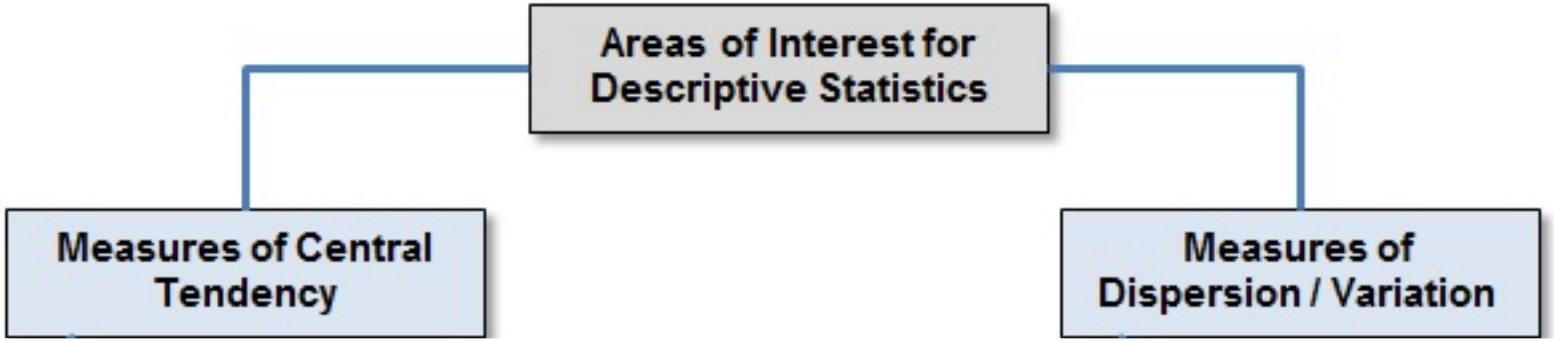
The influence of missing data on results

AuSCR
Australian Stroke Clinical Registry

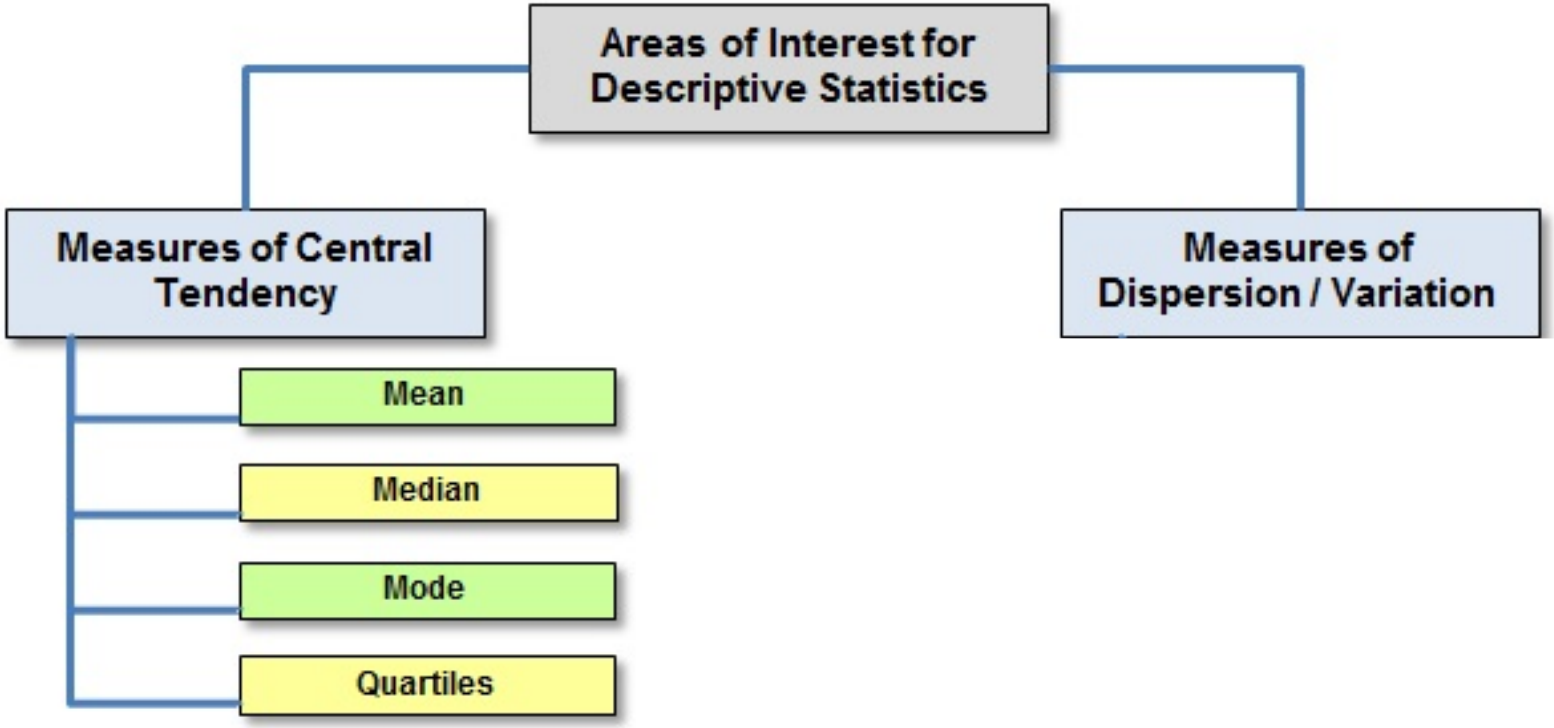
- Quality of clinical documentation
- Issues of missing data

	Missing data	Hospital adherence	Hospital adherence
		- missing	+ missing
Stroke unit care	0%	72%	72%
tPA for ischaemic strokes	0%	3%	3%
Aspirin in less than 48 hours (excl ICH)	0%	83%	83%
Mobilised same day or day after	25%	90%	68%
Oral screen before medications	7%	52%	48%
Discharge care plan	0%	67%	67%
Antihypertensive medication on discharge	9%	79%	72%
Antithrombotic medication on discharge (excl ICH)	0%	90%	90%
Lipid lowering medication on discharge	4%	70%	64%
tPA provided within 60 minutes of arrival	0 %	0%	0%

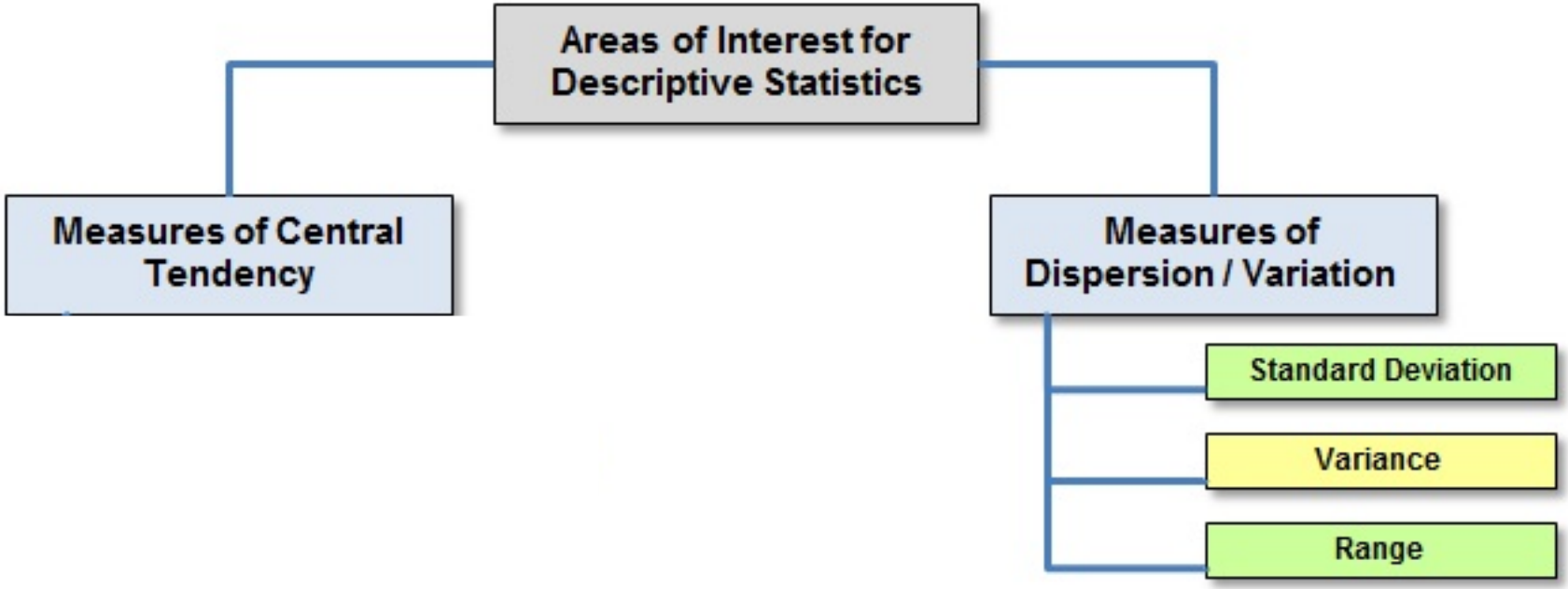
Descriptive statistics



Descriptive statistics

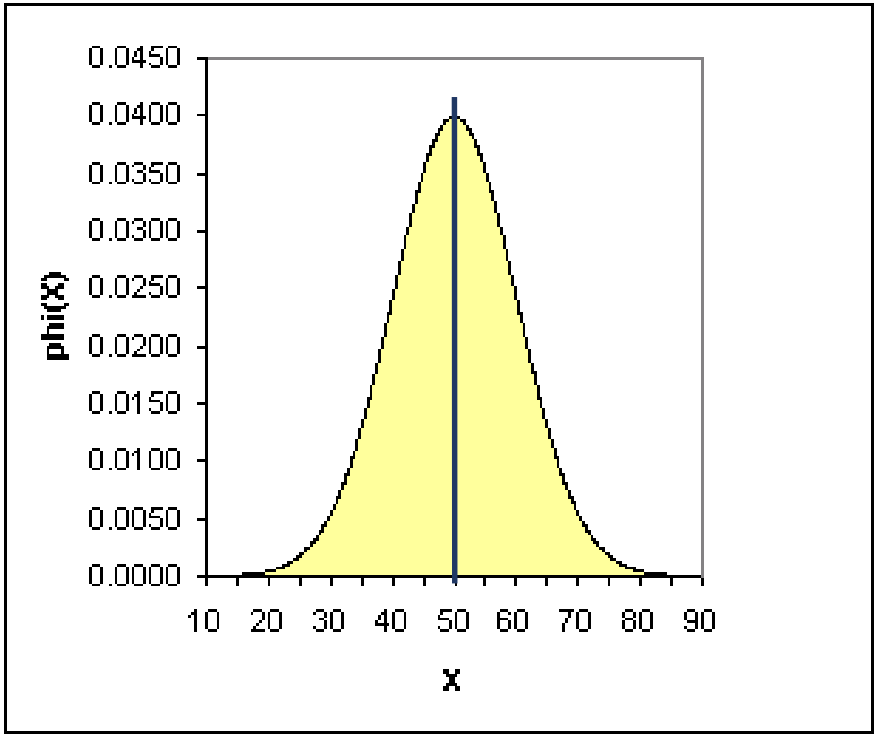


Descriptive statistics



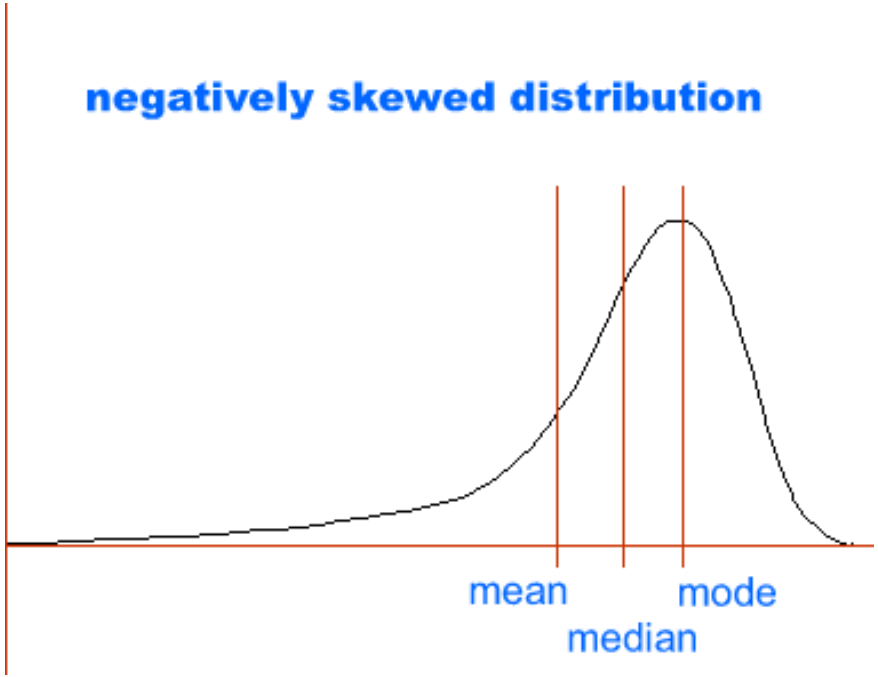
Parametric methods (normal distribution)

- Observations are independent
- Data are normally distributed
- Variances are equal across groups (e.g. bell shaped curve)
- Symmetrical around the mean
- When you calculate a mean, mode, and median they equal the same value



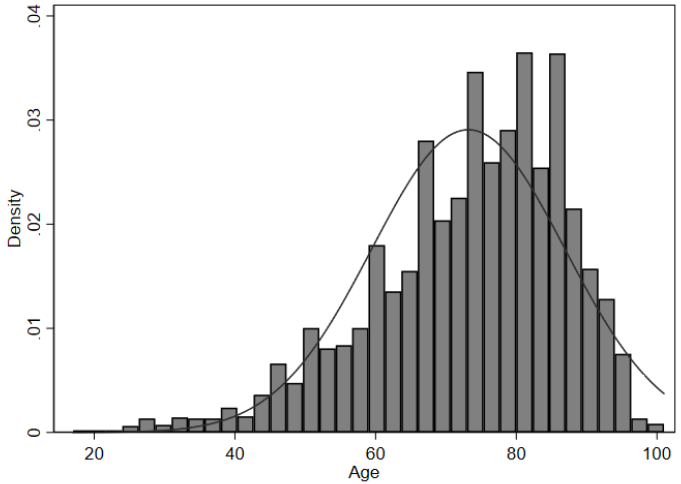
Non-parametric methods (non-normal distribution)

1. Non-parametric statistics do not assume any underlying distribution
2. Nonparametric statistics reduce data to an ***ordinal rank***, which reduces the impact or leverage of outliers
3. These statistics deal with the ***median*** rather than the mean

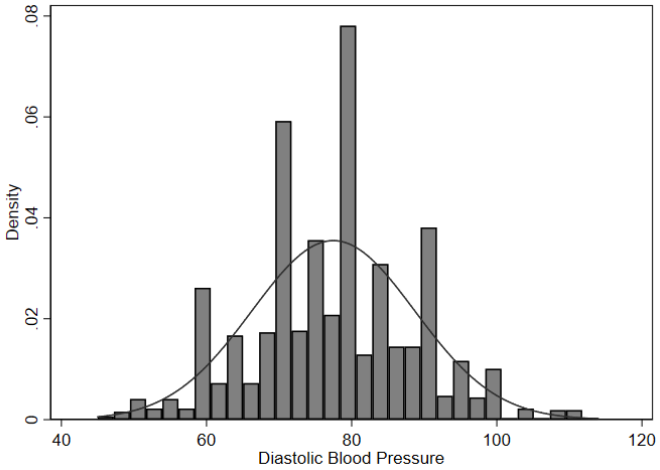


Distribution of data

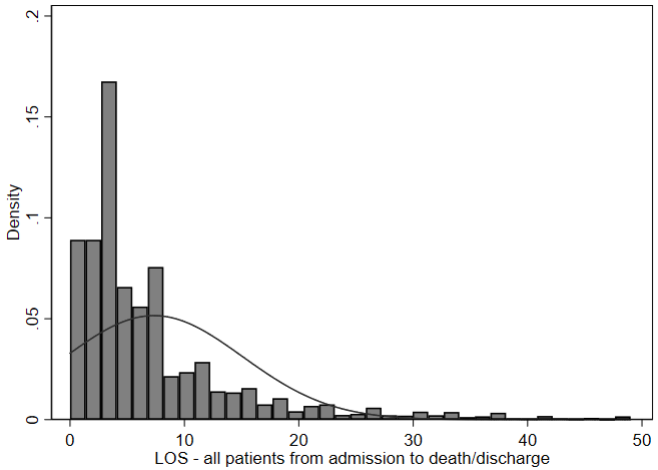
Negatively skewed



Symmetric



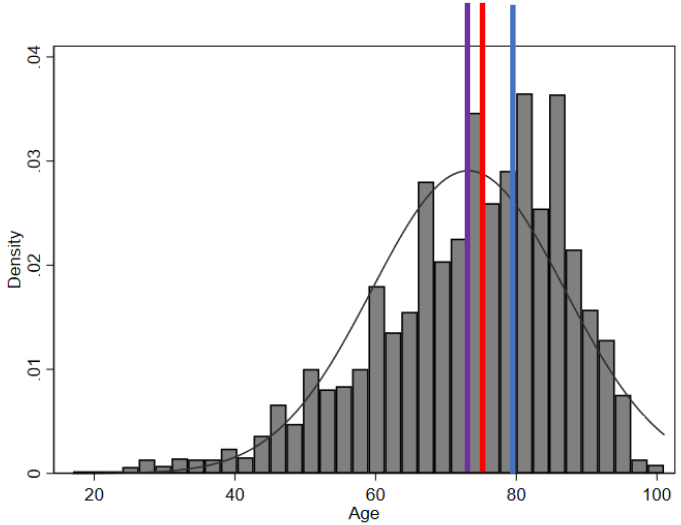
Positively skewed



Central Tendencies

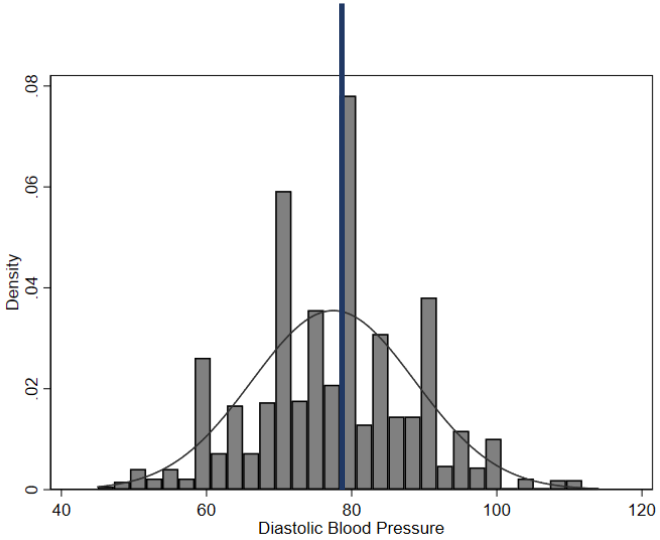
Negatively skewed

Mean ≤ Median ≤ Mode



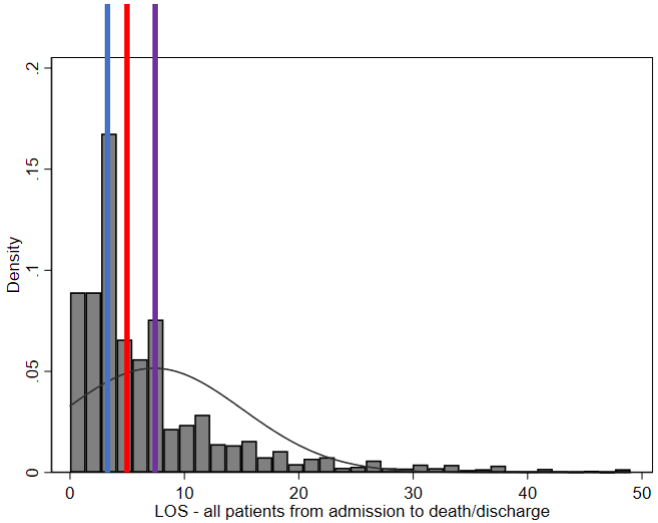
Symmetric

Mode = Mean = Median



Positively skewed

Mean ≥ Median ≥ Mode

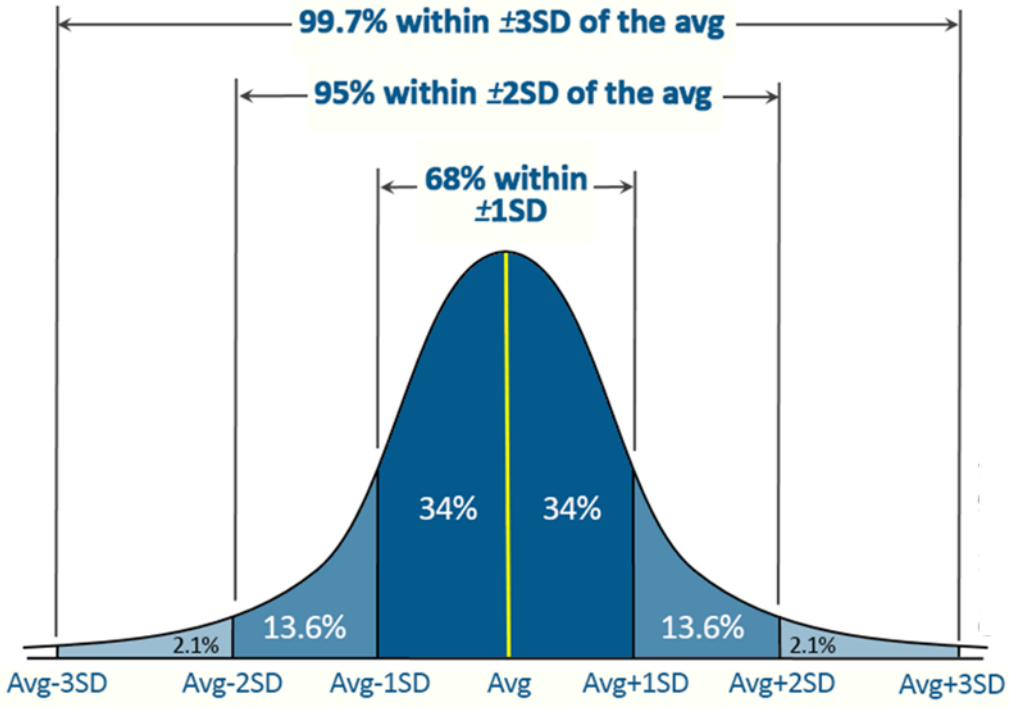


Mode: most frequently occurring value in a set of observations

Median: middle value that divides observations into two equal part (50th percentile)

Mean: average value of the observations (sum of observations/total observations)

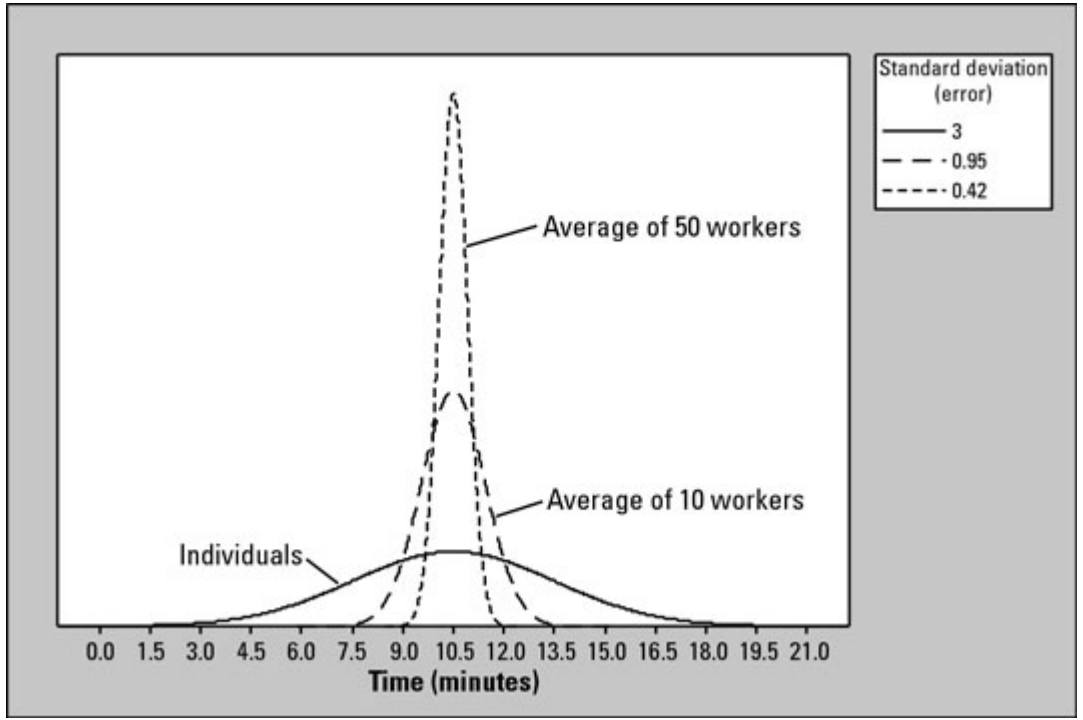
Standard deviation (SD)



Standard deviation tells you 'the spread of the data'
Used to describe how far each observed value is from the **mean**

Standard deviation

- Influenced by the sample size
- Standard deviation decreases with increasing sample size



Skewed data (non-parametric distributions)

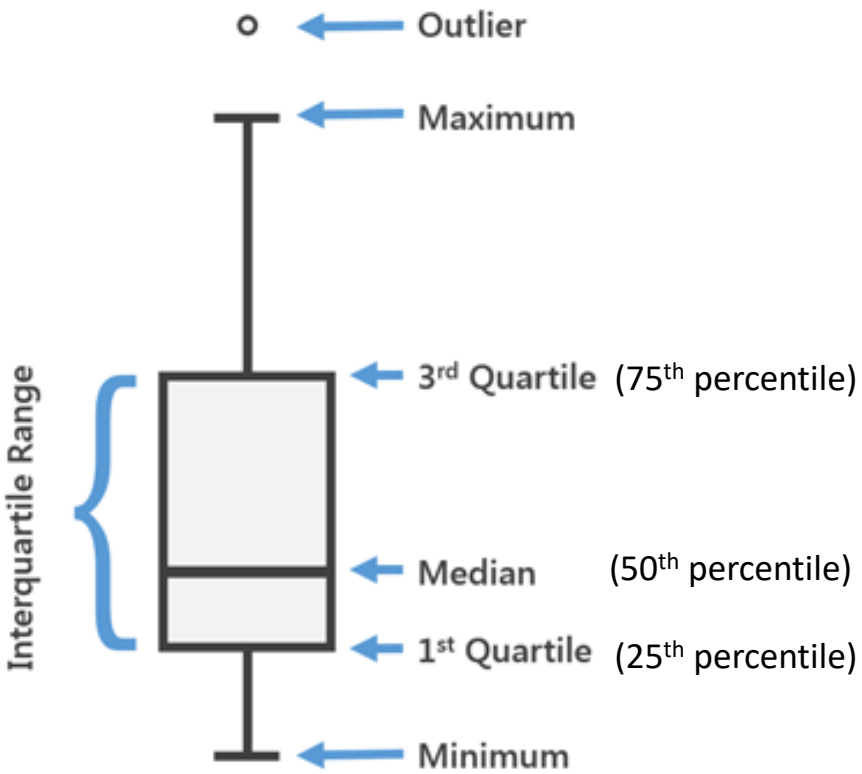
MEDIAN: This number has the property that it divides a set of observed values in two equal halves

- To calculate the **median**, the data should be arranged in order from smallest to biggest
- **Median:** the number that is halfway into the set
- If there is an even number of items in the dataset, then the **median** is found by taking the mean (average) of the two middlemost numbers
- ***Interpretation***: equal probability of falling above or below it

		1	1	
		2	2	
		3	3	
		4	4	
		5	5	
MIDDLE VALUE		6	10	
		7	11	
		8	12	
		9	20	
		10	24	
		11	25	
Median		6	10	
Mean		6	11	
SUM		66	117	
NUMBER OF VALUES		11	11	
MEAN	66/11		6 117/11	11

Interquartile range

- Reflects variability among middle 50% of observations
- IQR = Q3 –Q1
- Not affected by extreme values
- Used for asymmetric/non-parametric data

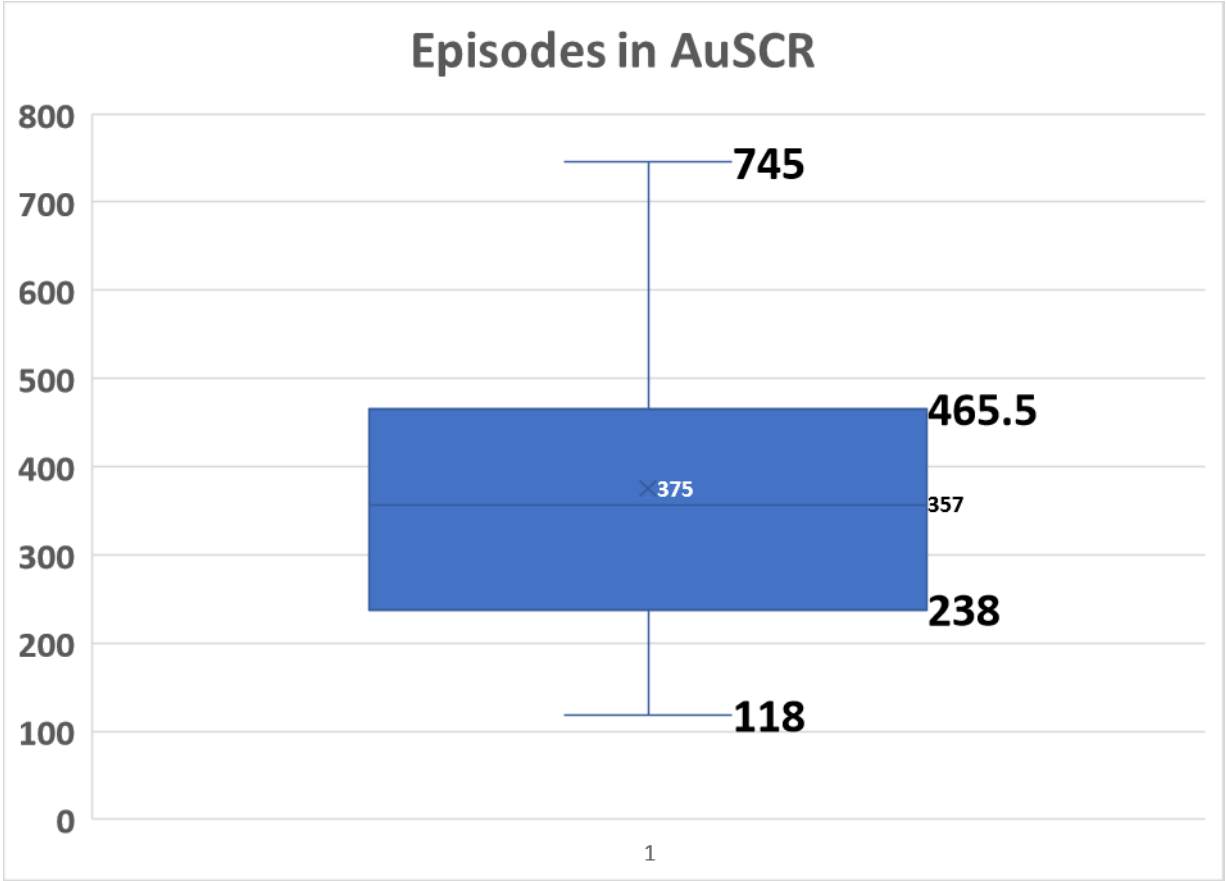


Example calculation of means from sample data (clusters) as part of national monitoring of hospitals

Hospital ID	Episodes in the AuSCR (n)	Episodes in hospital records not in the AuSCR (n)	Case ascertainment 2017
3	357	240	67%
5	346	110	32%
12	118	52	44%
13	130	80	61%
14	487	287	59%
15	745	663	89%
20	444	404	91%
MEAN or AVERAGE	375	262	63%
Min	118	52	32%
Max	745	663	91%
Standard deviation	217	217	22%

MONASH University

Mean or Median



MONASH University

MEAN or AVERAGE	375	262	63%
Min	118	52	32%
Max	745	663	91%
Standard deviation	217	217	22%
Q1 (25%)	**238**	**95**	**52%**
Q3 (75%)	**466**	**346**	**78%**
IQR = Q3 -Q1	**228**	**251**	**27%**

The interquartile range is the difference between the upper and lower quartile. IQR Q3 – Q1

Statistics

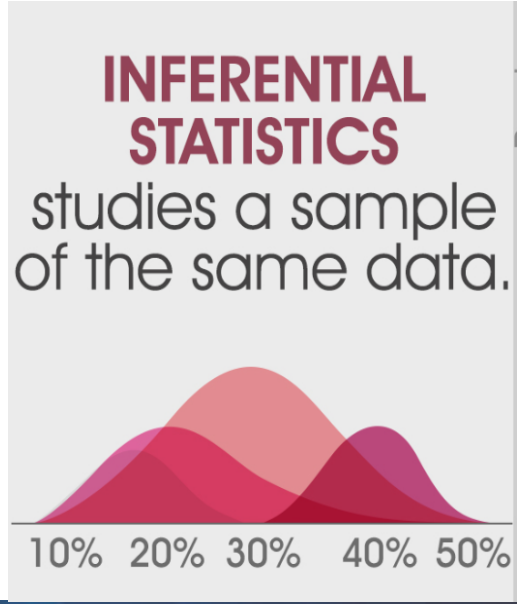
Descriptive:

collecting, organising, summarising, analysing and presenting data.

Inferential:

making inferences, hypothesis testing.

Determining relationship and making prediction.



The dreaded p-value = probability value

- p-value helps to determine the significance of your hypothesis testing
  - Null hypothesis: no true difference between groups
  - Alternate hypothesis: actual difference not attributed to chance

- Values between 0 and 1
- <0.05 generally used

What does my p-value mean?

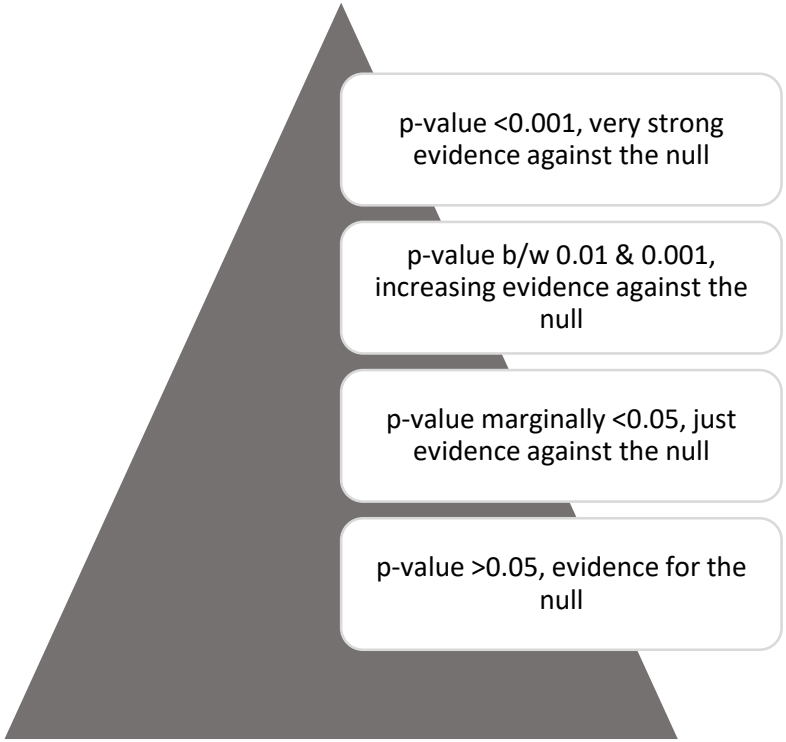
Is there is a difference in the proportion of females and males who receive tPA?

Null hypothesis

No difference in gender for % who receive tPA

Alternate hypothesis

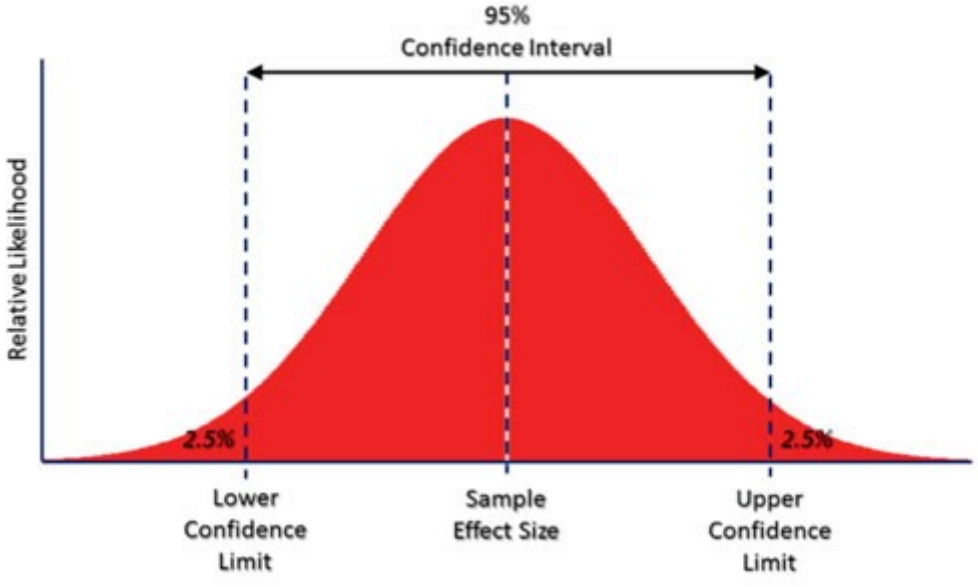
There is difference in the % of males and females who receive tPA



Confidence Intervals

- Indicates a range of values that's likely to encompass the 'true' value of the population parameter

- Provides an understanding of how much faith we can have in our sample estimates

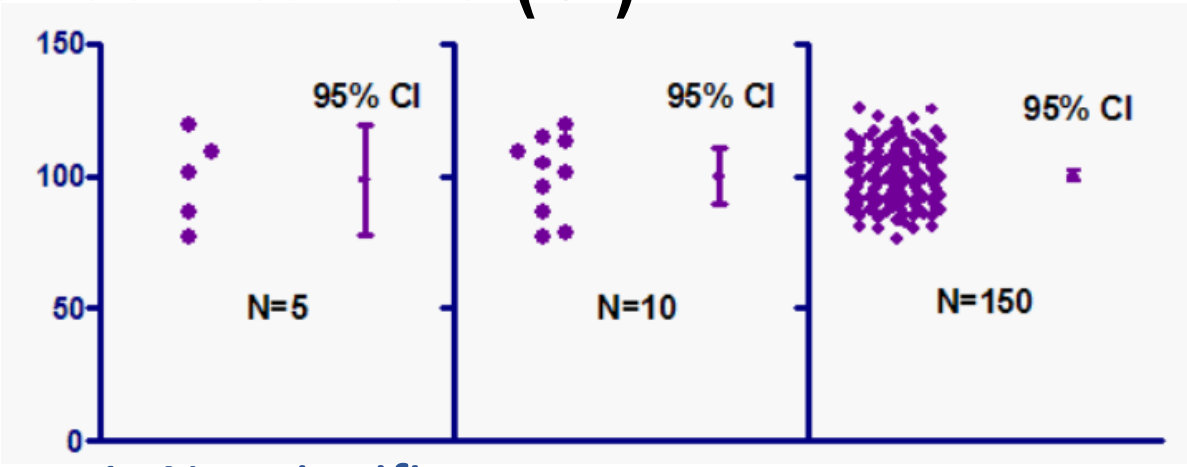


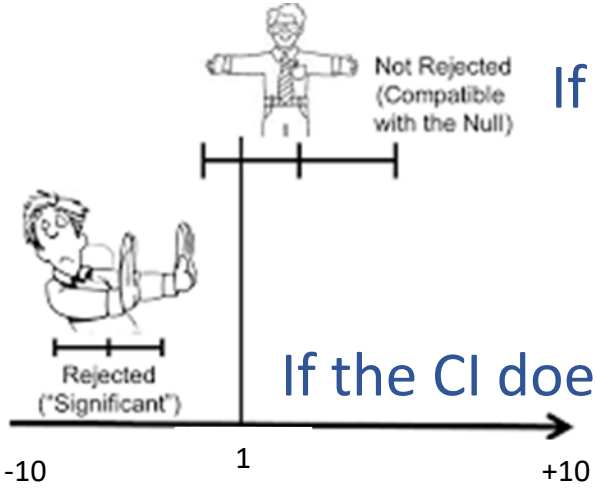
95% confidence interval for mean SBP for patients with acute stroke in Australia is (115, 132).

We can say with 95% confidence that the mean SBP for patients with acute stroke is between 115mmHg and 132mmHg

How to interpret confidence intervals (CI)

- Provides an idea of the sample size
- Indicates significance, but also direction of effect





If the CI crosses 1- Not significant

If the CI does not cross 1- Significant

Used with different estimates e.g. odds ratios

Our example

Is it true that there is a difference in the proportion of females and males who receive tPA?

% received tPA
Female- 9%
Male- 10%

p-value = 0.43
95% CI 0.87-1.38

What does it mean?

Null hypothesis
No difference in a set of

Alternate hypothesis
There is difference in the % of males and females who receive tPA

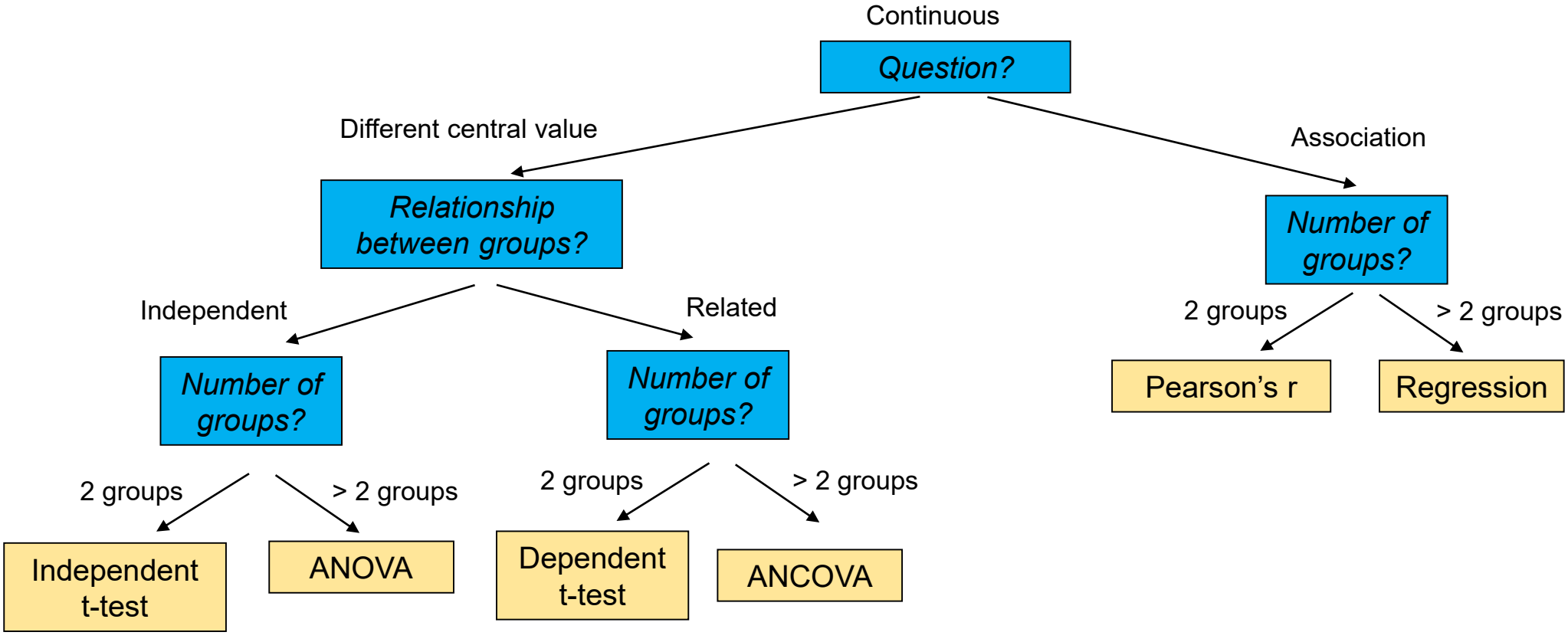
Still with me?

Shall we push on or stop for questions?

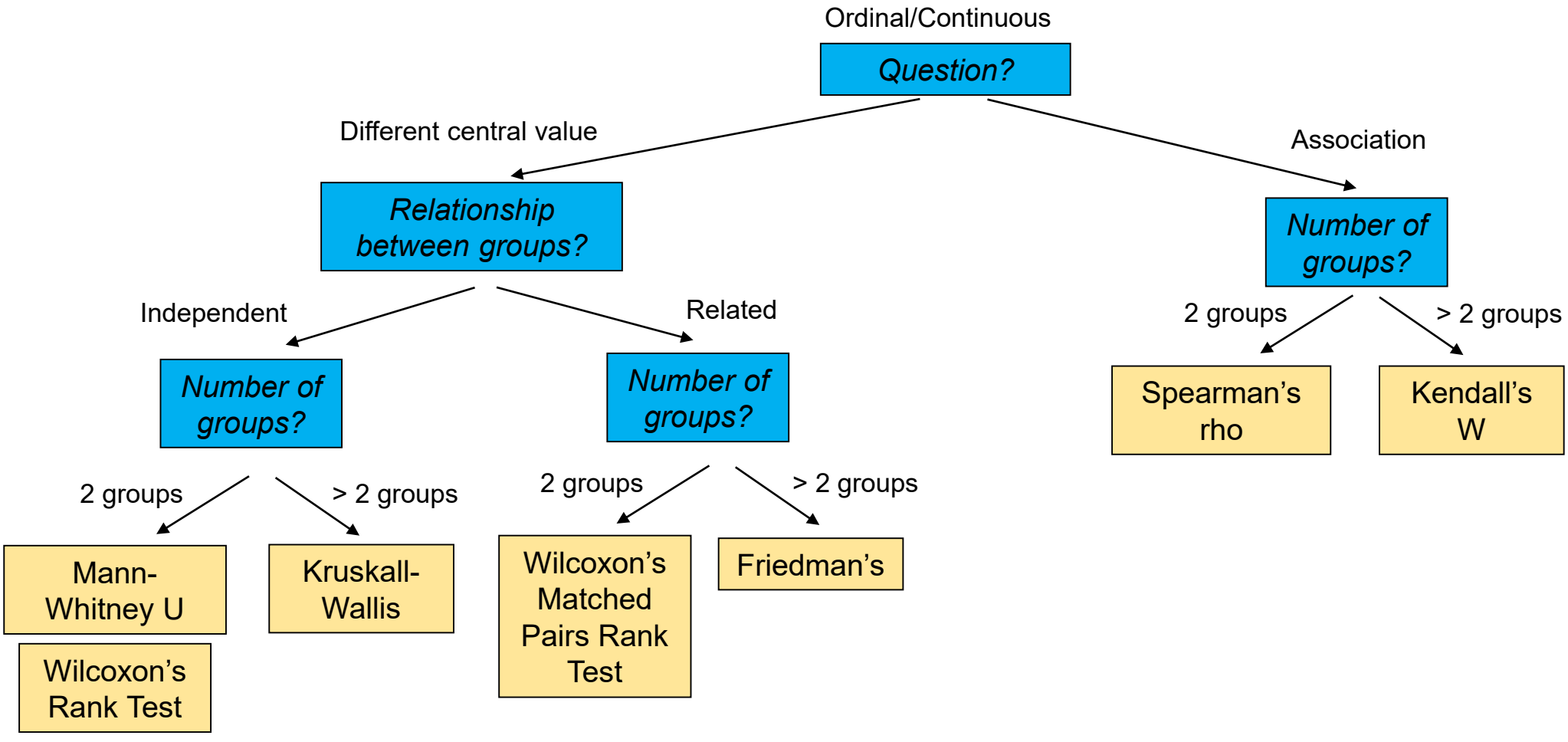
Other aspects to consider for a data analysis

- **How many groups?**
  - Two groups or more than two groups
- **Is there a relationship between the groups?**
  - Repeated measure (related) or independent

Parametric choices



Non-parametric choices



Other statistics you may also see in publications

Relative Risk (RR): is a ratio of the specified outcome in the exposed group compared with the control group. Ratio of two probabilities.

$$RR = \frac{\text{Risk of event in the Treatment group}}{\text{Risk of event in the Control group}} = \frac{a/(a+b)}{c/(c+d)}$$

Odds ratio (OR): Odds of an event occurring based on whether the patient has been exposed/treated.

In logistic regression analyses the odds ratio represents the constant effect of a predictor X, on the likelihood that one outcome will occur. Single summary score of the effect.

$$OR = \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}} = \frac{a/b}{c/d} = ad/bc$$

Relative Risk Reduction (RRR)

Calculated as 1 – RR. Therefore, a RR of 0.8 means a RRR of 20% which is a 20% reduction in the relative risk of the specified outcome in the exposed group compared with the control group.

Absolute Risk Reduction (ARR)

Net difference between rates of disease among exposed (intervention) and unexposed (control) groups.

Number Needed to Treat

This is the number of patients who need to be treated to prevent one case of disease and is estimated using the following formula **1/ARR**.

Interpreting Odds Ratios

- An OR of 1 indicates that odds of cases/disease are the same in the exposed/treated and unexposed/non treated (control) groups
- If the **OR > 1**, odds of cases/disease are greater in the treated group compared with the control group
  - OR of 1.5 for independence = Compared to no treatment, those treated have a 50% greater odds of independence
- If the **OR < 1**, the odds of cases in the treated group are less than the odds of cases in the control group
  - OR of 0.6 for death = Compared to no treatment, the odds of death are 40% less for those treated

Example

- Odds of an event occurring based on whether the patient has been exposed/treated.
- The OR can approximate the Relative Risk for rare outcome events

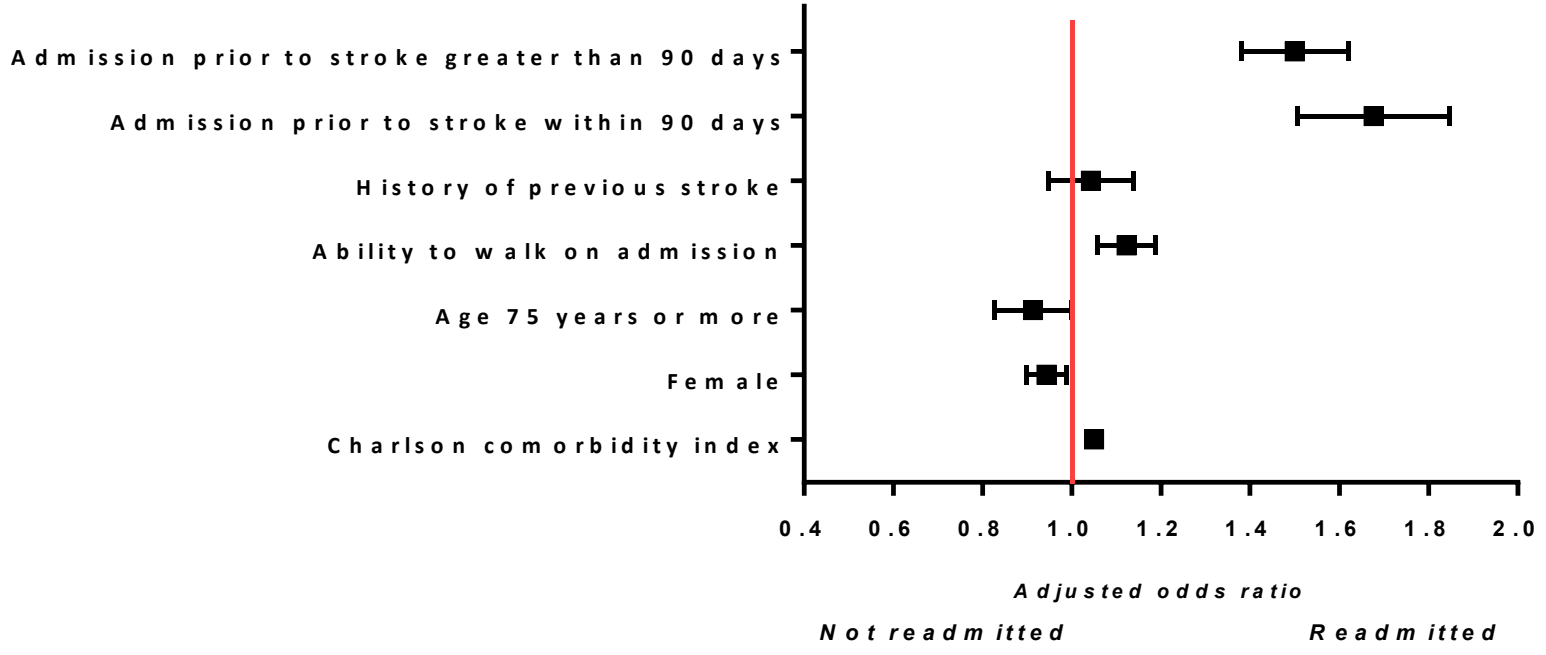
	Death	No Death	TOTAL N
Treatment	**8**	**992**	**1000**
Control	**10**	**990**	**1000**
TOTAL N	**18**	**1982**	

The **OR** = (8*990) / (10*992) = 0.8 making the odds of better outcome 1-OR = 20%.

Absolute risk reduction is:
(10/1000) -(8/1000)= 0.01-0.008= 0.002 or 0.2%
(because you multiply answer by 100).

NNT is 1/0.002=500 patients.

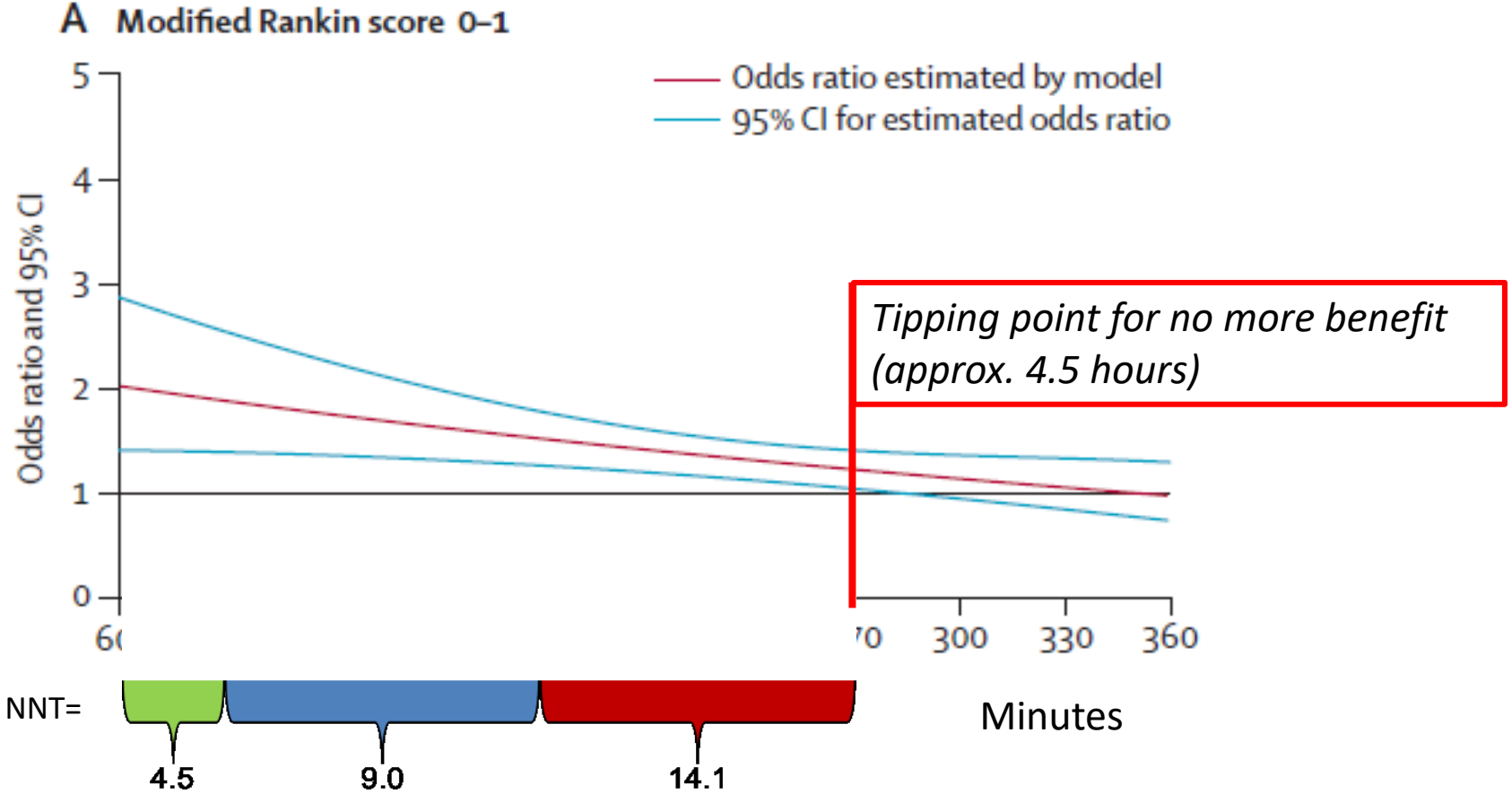
Understanding an odds ratio and confidence interval



Odds ratio with 95% confidence interval

- 1.0 no effect/ difference (if confidence interval crosses 1.0) e.g. history of previous stroke
- <1= factor associated with not being readmitted e.g. age
- >1= factor associated with being readmitted e.g. admission prior to stroke

Effectiveness of intravenous thrombolysis



Lees KR, et al. *Lancet*. 2010;375:1695-1703

Example for Endovascular clot retrieval (ECR)

The aggregate clinical trial data strongly favor ECR over medical management for anterior circulation LVO.

The number needed to treat (NNT) is 2.6 for one additional patient to achieve improved functional outcome, defined as improvement by at least one level on the modified Rankin Scale (mRS) at 90 days (adjusted odds ratio [aOR] = 2.49, 95% confidence interval [CI] = 1.79–3.53).

Additionally, for every 5 patients treated with ECR, one additional patient achieved functional independence (mRS 0–2) at 90 days (46%vs. 26.5%, absolute risk difference = 19.5%, aOR = 2.71, 95% CI = 2.07–3.55)

NNT = 1/0.195 = 5.12

https://www.thennt.com/nnt/early-endovascular-thrombectomy-large-vessel-ischemic-stroke-reduces-disability-90-days/

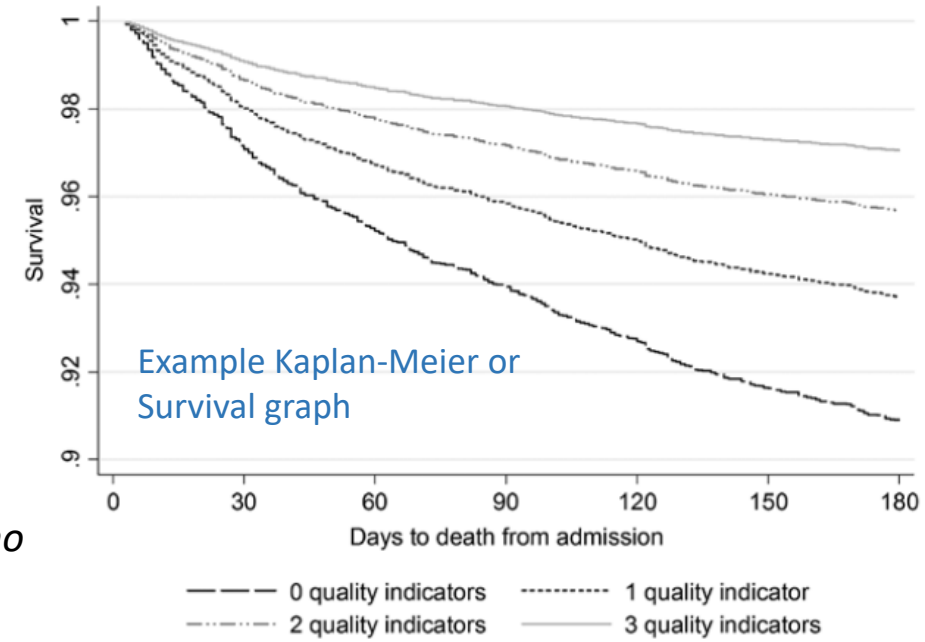
Hazard Ratios

- The hazard ratio is a measure of an effect of an intervention (i.e. SU care) on an outcome (i.e. survival at 3 months).

$$\text{HR} = \frac{\text{hazard in the intervention}}{\text{hazard in the control}}$$

"Hazard" refers to the probability that an individual, under observation in a clinical trial or observational study at time t, has an event at that time
If Hazard ratio = 1 means no risk reduction compared to the control

"Those who received all 3 indicators compared to those who received none, had a 70% reduced hazard of death at 180 days (HR 0.30; 95% CI 0.18-0.47)"



Example Kaplan-Meier or Survival graph

Cadilhac et al, *Stroke*, 2017

MONASH University

Univariable analyses or 'Descriptive analysis'

- Simple or univariable analyses: only two variables (dependent and independent variable) e.g. difference in access to stroke unit by gender

Discharged	Weekend	Weekday	Unadjusted odds ratio (95% CI)
Treated in a stroke unit	69%	81%	0.53 (0.48, 0.58)
Discharged on antihypertensive medication	65%	71%	0.74 (0.68, 0.81)
Discharge with a care plan	47%	53%	0.78 (0.70, 0.86)
Swallow screen or assessment	63%	86%	0.28 (0.25, 0.33)

Multivariable analyses

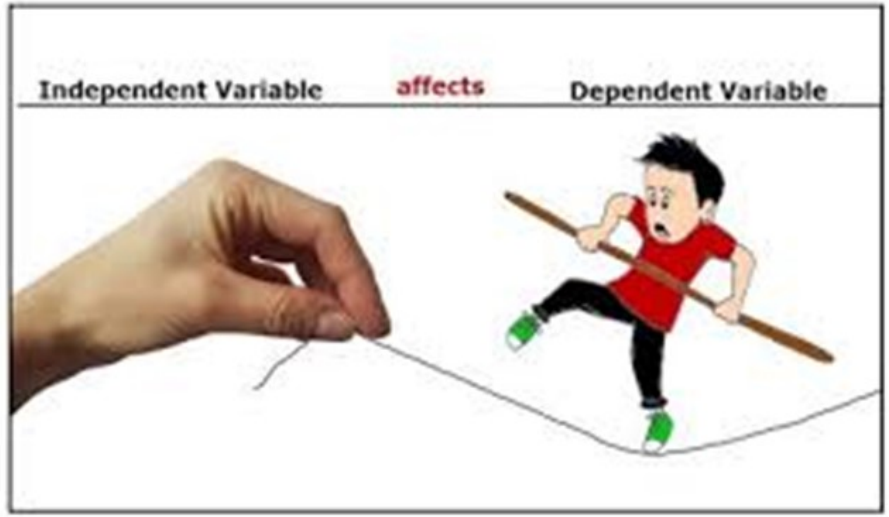
- Multivariable statistical models used in hypothesis testing
- One dependent variable and 2+ independent variables

- Used to find patterns and relationships between several variables simultaneously

- Used particularly with outcomes
  - Length of stay can be influenced by patient age, stroke severity and type

- Limit chance of ‘false positive’-
  falsely conclude a difference exists

Variables used in multivariable models



Dependent variable

Outcome of interest

- Mortality
- Quality of life
- Readmission
- Process of Care

Independent variables: *confounders*

- Treatment in a stroke unit
- Stroke type
- Age, sex, etc

Confounding variables: may compete with the exposure of interest (eg, treatment) in explaining the outcome of a study. The amount of association "above and beyond" what can be explained by **confounding** factors provides a more appropriate estimate of the true association due to the exposure.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

Outcomes

Risk adjusted mortality between hospitals

Adjusted for age, sex, stroke type, ability to walk on admission, and past history stroke

3 SD

2 SD

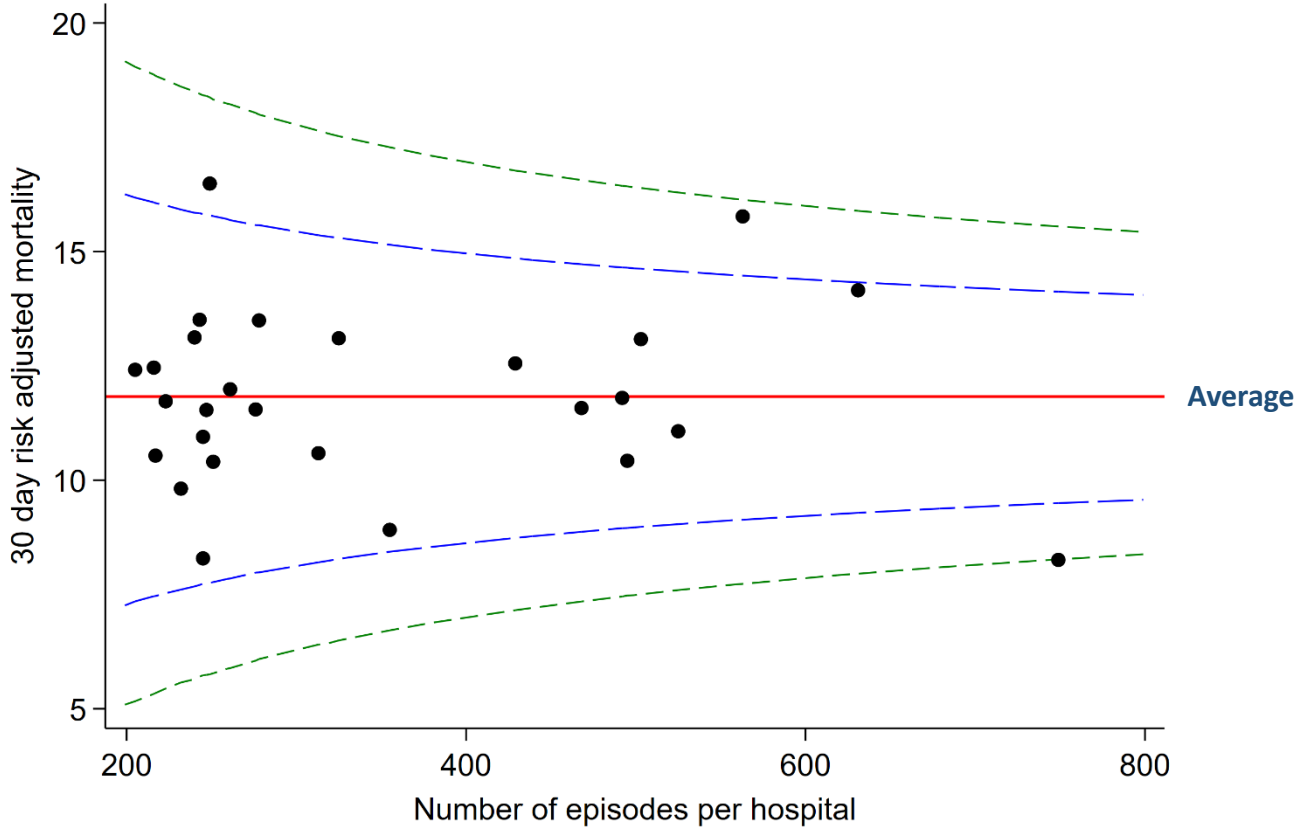


Table 4 Outcomes of care in hospitals admitting at least 100 patients in a year by stroke unit status

Outcomes Reference: stroke unit hospital versus no stroke unit hospital	SU hospital n (%) *N* = 2,481	Non-SU hospital n (%) *N* = 417	Self-reported SU status *N* = 81 Odds ratio (95% CI)[a]	Meet all SU criteria *N* = 59 Odds ratio (95% CI)[a]
Modified Rankin Score (mRS) on discharge 0–2 (i.e., none to slight disability) [29]	885 (36)	106 (25)*	1.09 (0.65, 1.84)	0.94 (0.59, 1.52)
Stroke progression (including hemorrhagic transformation)	256 (10)	34 (8)	1.28 (0.67, 2.47)	1.14 (0.64, 2.04)
New stroke (recurrent event in hospital)	42 (2)	35 (8)*	0.20 (0.06, 0.61)*	0.25 (0.08, 0.74)*
Discharge destination[b]				
Discharged home	985 (45)	155 (45)	0.84 (0.48, 1.50)	0.88 (0.52, 1.50)
Discharged to inpatient rehabilitation	859 (35)*	98 (24)	1.23 (0.77, 1.96)	1.18 (0.77, 1.80)
Discharged to an aged care facility	219 (10)	51 (15)*	0.72 (0.36, 1.46)	0.70 (0.37, 1.34)
Died in hospital	291 (12)	75 (18)*	0.57 (0.33, 1.00)**	0.51 (0.31, 0.83)*
Died or discharged to aged care facility	510 (21)	126 (30)*	0.61 (0.36, 1.02)	0.58 (0.36, 0.92)*

[a]Each outcome was adjusted for hospital stroke unit status (self-reported or if all the Acute Stroke Services Framework stroke unit criteria were met[9]), age, sex, stroke severity variables (e.g., unable to walk on admission), independent prior to stroke and type of stroke (ischemic, intracerebral hemorrhage or unknown type)
[b]Excludes discharge destination noted as a statistical discharge (11% of patients); *$p < 0.05$; **$p < 0.07$

Cadilhac et al, *BMC Health Services Research*, 2017

MONASH University

Self-assessment

PollEv.com/tarapurvis583

MONASH University

Self-assessment questions

1. A visual analogue pain scale (0-10) is an example of what type of variable

   (nominal, ordinal, continuous, unsure)

2. What is the best measure to describe the age of your cohort if it is not normally distributed

   (mean, percentage, mode, median, unsure)

3. A subgroup that is representative of a population such as ischaemic stroke is:

   (a category, data, a sample, unsure)

4. A study reports a p value of 0.07. Usually, this means the result is:

   (significant, not significant, unable to be determined from this information, unsure)

5. Odds of death in treatment versus control group was reported as OR 1.72, 95% CI 0.9, 1.93. What does this mean?

   - Patients in the treatment group were more likely to die compared to the control - the result was significant
   - Patients in the control group were more likely to die compared to the treatment group - the result was not significant,
   - Patients in the treatment group were more likely to die compared to those in the control group - the result was significant, unsure)

Summary

- Covered a range of introductory concepts
  - Understanding variance in samples and summary statistics
- Provided information to assist in understanding data
- Pointers for critical appraisal of the literature related to methods and quantitative data

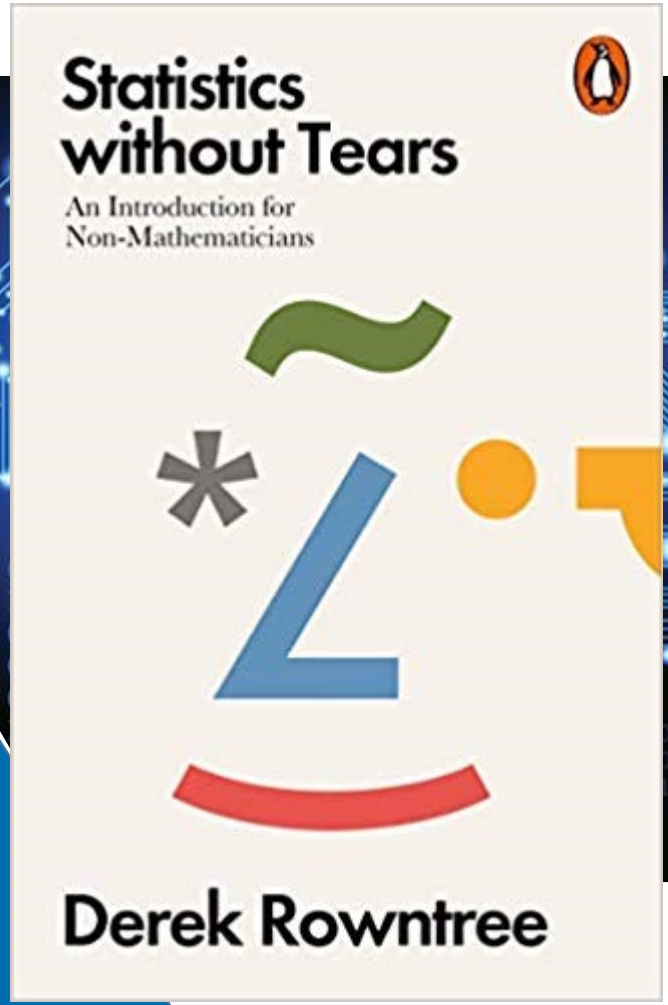
Acknowledgements

Adaption of some slides from
Monique Kilkenny and Tara Purvis

Further information

dominique.cadilhac@monash.edu

tara.purvis@monash.edu



MONASH University

Other useful sources of information

https://www.theanalysisfactor.com/the-difference-between-relative-risk-and-odds-ratios/

https://www.theanalysisfactor.com/why-use-odds-ratios/

http://www.ebm.med.ualberta.ca/TherapyCalc.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/