STATISTICS

Fundamentals & Clinical Application

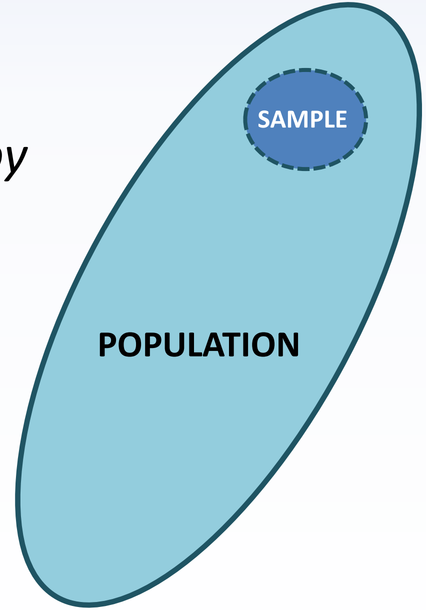
Research & Innovation Committee

James Stockley

ARTP

Introduction

- Statistics are fundamental to medical science
- They are concerned with **estimation**;
  - *We estimate what we think is true of a* ***population*** *by studying a representative* ***sample***
- The challenge is determining if what we observe is **real** or **artefact** due to variation



Hypotheses

- A supposition based on limited knowledge… a **starting point** for further investigation

- In terms of statistics;
  - Null hypothesis (**H_0**) = no effect

    … assumed true unless there is strong evidence to the contrary

  - Alternative hypothesis (**H_A** or **H_1**) = a true effect

Hypotheses

- For example;

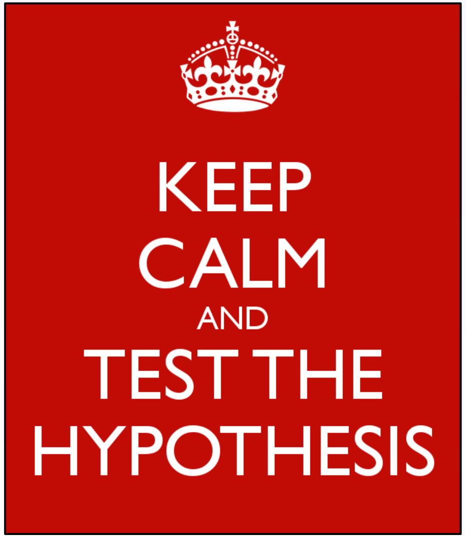
  "I hypothesise that chronic e-cigarette vaping causes emphysema"

  H_0 = e-cigs do not cause emphysema

  H_A = e-cigs cause emphysema

Hypotheses

- H_0 and H_A must be defined **before** statistical tests are selected
- It is the **question** that dictates the methodology and which statistical tests are appropriate



POWER!

"The probability that a test will detect an effect when there is an effect to be detected"

https://www.cio.com

Power

- The probability that what we are observing is true
- Dependent on the size of the **effect** and the size of the **sample**
- A “best guess” based on current knowledge of outcome measures from **previous research** or own **pilot data**

Power

- Power calculations commonly yield a sample size that will provide **80% probability** the observations are true

- Designed to minimise the risk of Type I and Type II errors

	Do not Reject H_0	**Reject H_0**
H_0 is True	Correct Decision	Incorrect Decision “Type I Error” (β)
H_0 is False	Incorrect Decision “Type II Error” (α)	Correct Decision

Example website for calculating power: http://www.stat.ubc.ca/~rollin/stats/ssize/

Power

Stating an effect when there isn't one
"over-powered"

- Designed to minimise the risk of Type I and Type II errors

	Do not Reject H_0	Reject H_0
H_0 is True	Correct Decision	Incorrect Decision "Type I Error" (β)
H_0 is False	Incorrect Decision "Type II Error" (α)	Correct Decision

Example website for calculating power: http://www.stat.ubc.ca/~rollin/stats/ssize/

Power

Stating no effect when there is one
"under-powered"

- Designed to minimise the risk of Type I and Type II errors

	Do not Reject H_0	Reject H_0
H_0 is True	Correct Decision	Incorrect Decision "Type I Error" (β)
H_0 is False	Incorrect Decision "Type II Error" (α)	Correct Decision

Example website for calculating power: http://www.stat.ubc.ca/~rollin/stats/ssize/

Types of Data

1. Categorical

i. Nominal – *mutually exclusive groups that cannot be ordered (e.g. sex, race)*

ii. Ordinal – *groups ranked in order of magnitude, difference between groups <u>not</u> identical (e.g. GOLD classification, BORG score*)*

2. Numerical

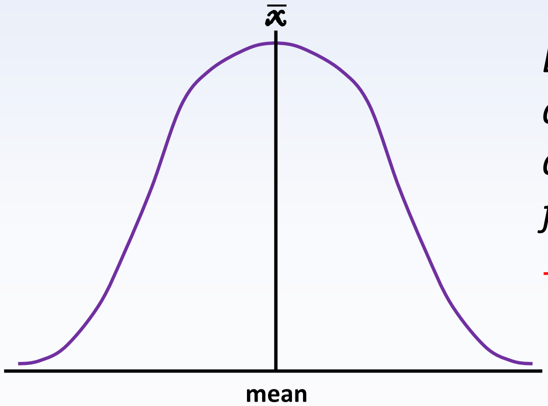
i. Discrete – *groups ranked in order of magnitude, difference between groups <u>is</u> identical i.e. "**countable**" (e.g. number of exacerbations per year)*

ii. Continuous – *any value within a range i.e. "**measurable**" (e.g. height, FEV_1)*

Distribution

- **Normal**

  a.k.a. "Gaussian"

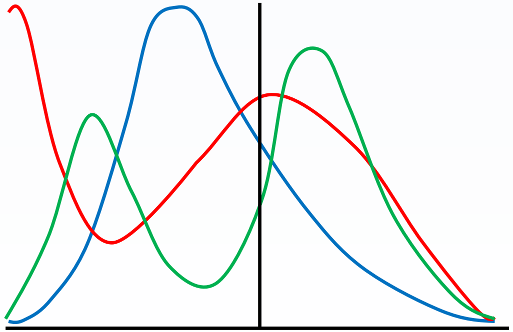


Data are distributed symmetrically around the mean ($\bar{x}$), with data around the mean occurring more frequently

→ ***Parametric Tests***

- **Not normal…** *anything else*

  → ***Non-Parametric Tests***



Determining Distribution

- **Statistical**
  - Shapiro-Wilk test: *Small samples ($n < 50$)*
  - Kolmogorov-Smirnov test: *Medium-large samples ($n \geq 50$)*

- **Visual**
  - Stem-and-Leaf plot / Histogram
  - Quantile-Quantile (Q-Q) plot

Determining Distribution

- **Stem & Leaf Plot**

Data set 1: 4, 6, 12, 19, 20, 22, 24, 27, 31, 31, 31, 32, 36, 38, 39, 39, 44, 45, 47, 47, 48, 49, 49, 53, 55, 56, 56, 60, 61, 77.

0	4	6						
1	2	9						
2	0	2	4	7				
3	1	1	1	2	6	8	9	9
4	4	5	7	7	8	9	9	
5	3	5	6	6				
6	0	1						
7	7							

STEM

LEAF

http://whattest.lboro.ac.uk/glossary/normal%20distribution.pdf

Determining Distribution

- **Stem & Leaf Plot**

Data set 1: 4, 6, 12, 19, 20, 22, 24, 27, 31, 31, 31, 32, 36, 38, 39, 39, 44, 45, 47, 47, 48, 49, 49, 53, 55, 56, 56, 60, 61, 77.

0	4	6						
1	2	9						
2	0	2	4	7				
3	1	1	1	2	6	8	9	9
4	4	5	7	7	8	9	9	
5	3	5	6	6				
6	0	1						
7	7							

http://whattest.lboro.ac.uk/glossary/normal%20distribution.pdf

Determining Distribution

- **Histograms**

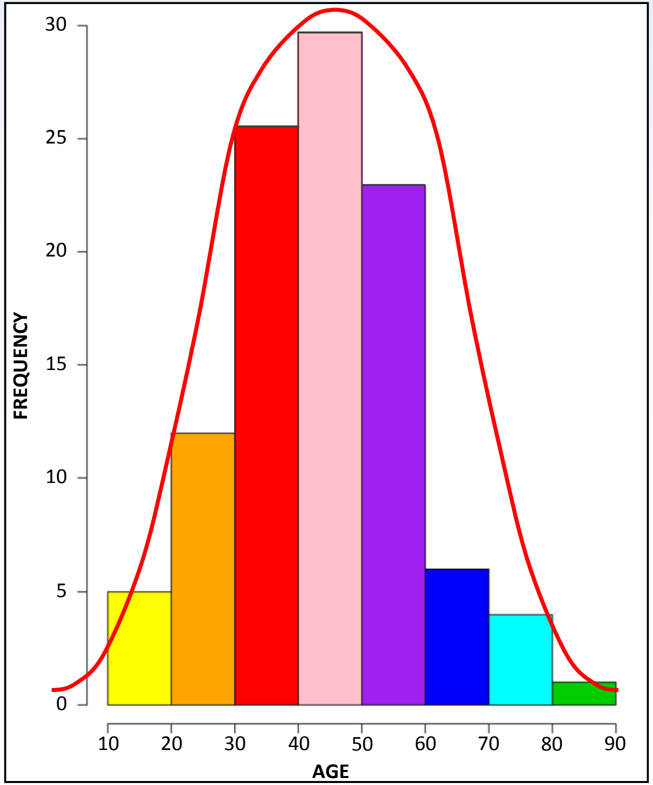
Data set 1: 18, 18, 19, 19, 20, 21, 24, 27, 29, 29, 29, 30, 30, 30, 30, 30, 30, 32, 32, 32, 32, 32, 32, 33, 34, 34, 34, 34, 35, 35, 36, 37, 37, 38, 38, 39, 39, 40, 40, 40, 41, 41, 41, 41, 41, 41, 41, 41, 43, 43, 43, 43, 43, 44, 44, 45, 45, 46, 48, 48, 48, 48, 48, 48, 48, 49, 50, 51, 52, 52, 52, 52, 52, 53, 53, 53, 53, 54, 55, 56, 57, 57, 57, 57, 57, 57, 57, 58, 59, 64, 66, 67, 67, 68, 69, 72, 76, 78, 79, 79.

Age	Frequency
$10 \leq x < 20$	4
$20 \leq x < 30$	7
$30 \leq x < 40$	26
$40 \leq x < 50$	29
$50 \leq x < 60$	23
$60 \leq x < 70$	6
$70 \leq x < 80$	5

http://whattest.lboro.ac.uk/glossary/normal%20distribution.pdf

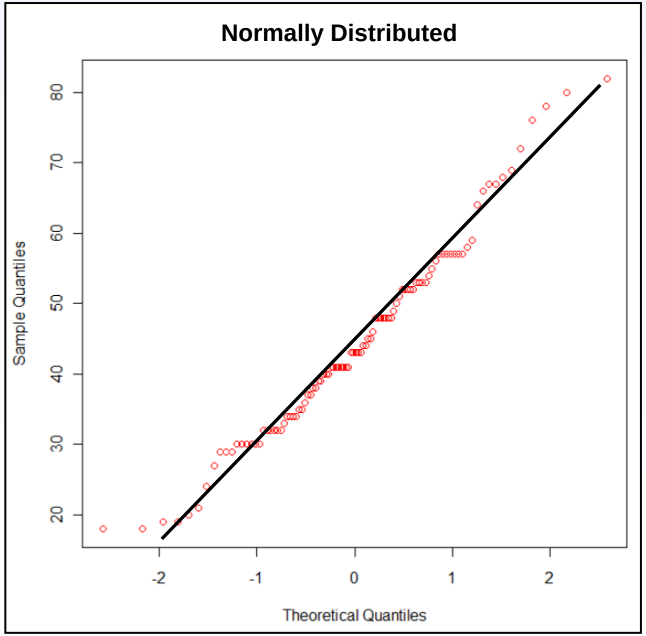
Determining Distribution

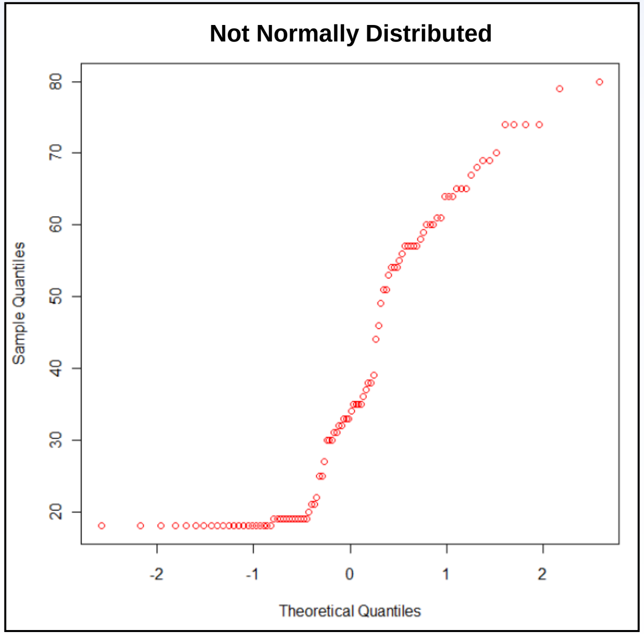
- **Histograms**



Determining Distribution

- **Q-Q Plot**





http://whattest.lboro.ac.uk/glossary/normal%20distribution.pdf

Statistical Significance

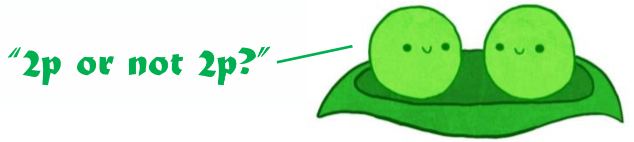
- **$p < 0.05$** is usually used as the threshold for statistical significance

- In other words, if $p < 0.05$, the probability that the observation happened by chance is minimal (<5%)

Fisher, R.A. (1925) Statistical Methods for Research Workers. Oliver & Boyd, Edinburgh

Statistical Significance

- If an effect is **predicted** to go one way, simply use **p** ("one tailed")
  - E.g. Effect of a novel treatment for sleep apnoea – AHI would be *expected* to reduce
- If it is **unknown** what direction the effect could go, use **2xp** ("two tailed")
  - E.g. Number of neutrophil surface chemoreceptors in COPD compared to health – could be greater *or* fewer



REGRESSION

https://www.von.gov.ng/

Linear Regression

- Determines if there is an **association** (correlation) between two sets of numerical data (variables)

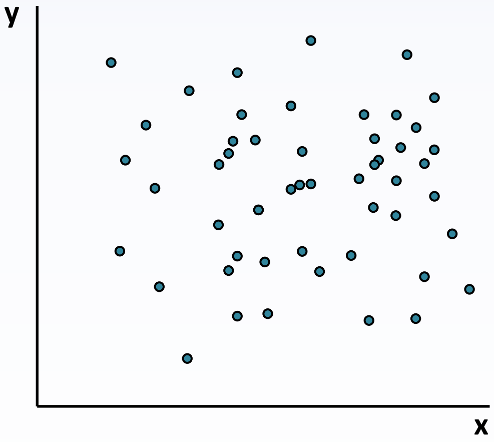
- Start by producing a **scatter plot** of *x* vs *y*;

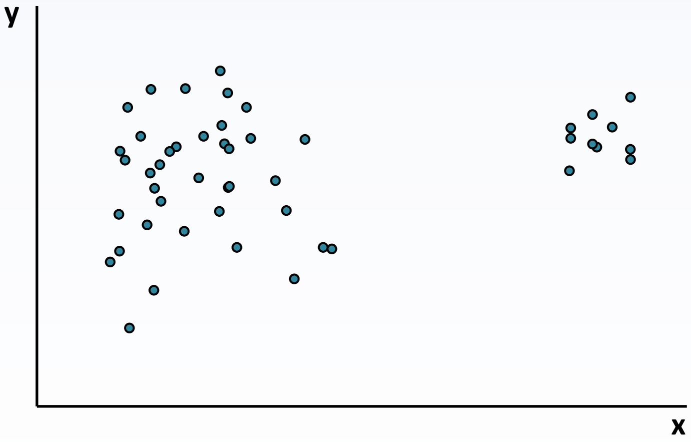
  x – abscissa… the *explanatory* variable
  y – ordinate… the *dependent* variable

- **Visually** assess the data first

Linear Regression

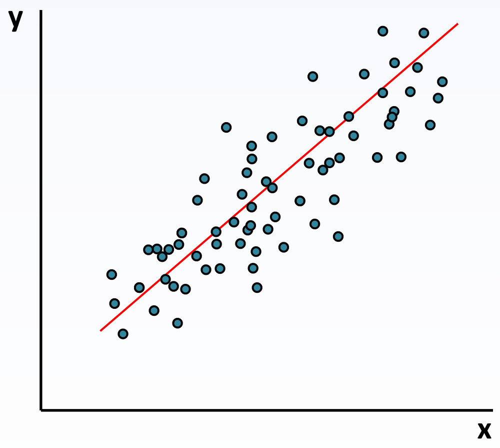
- No correlation;



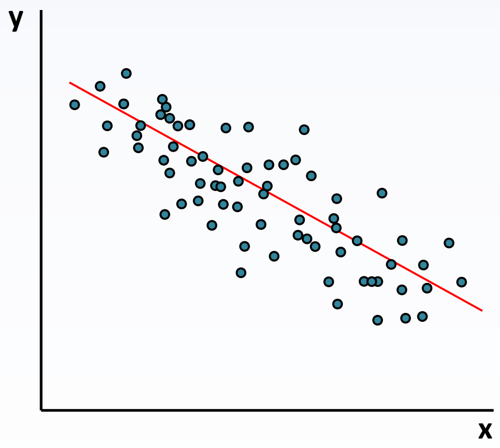


Linear Regression

- Positive Correlation;



- Negative Correlation;



Testing Correlation

- Correlation can be assessed statistically to determine **significance** and **strength** of the relationship;

  – **Parametric** - *Pearson's correlation*

  – **Non-parametric** - *Spearman Rank correlation*

r or r^2 ?

- **r** is the "**correlation coefficient**"

- It denotes the overall strength of the correlation;
  - **r = 1** is a perfect *positive* correlation
  - **r = 0** absolutely no correlation
  - **r = -1** is a perfect *negative* correlation

- The closer r is to +1 or -1, the greater is the strength of the association

r or r^2 ?

- r^2 is the "**coefficient of determination**"

- It measures the proportion of variation in **y** (e.g. AHI) that is explained by **x** (e.g. BMI)

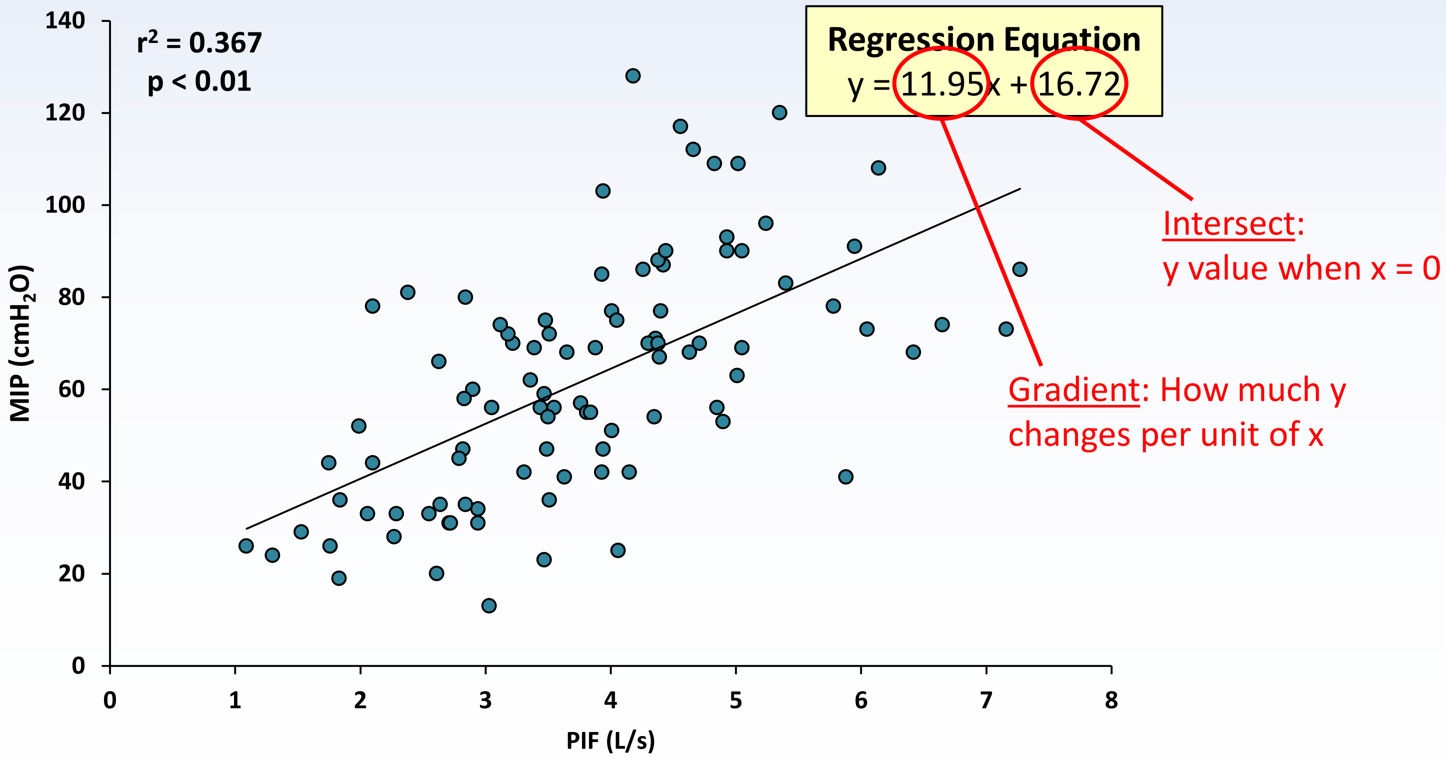
- It is more appropriate to quote r^2 when the research question concerns the **dependence of y on x**

Strength of r^2

r^2 Value	Strength of Relationship
< 0.3	None / very weak
0.3 – 0.5	Weak
0.5 – 0.7	Moderate
> 0.7	Strong

Moore D. S., Notz, W. I., & Fligner, M. A. (2013). The basic practice of statistics (6th ed.). New York, NY: W. H. Freeman and Company. Page 138

Regression Equation



Correlation NOT Causation

- In statistics, correlation does not imply causation
- Cause-and-effect cannot be legitimately deduced based solely on an association

Example

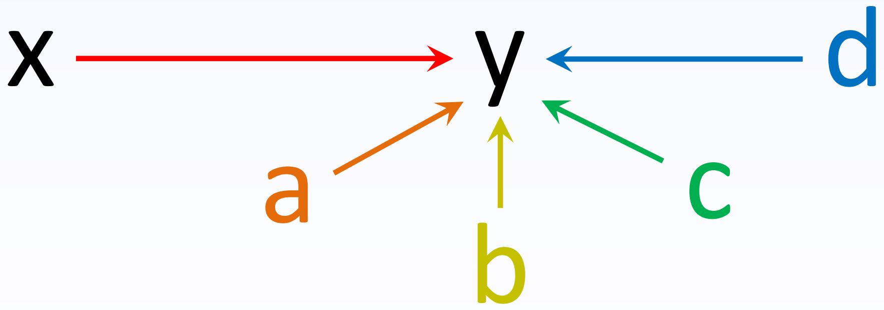
Women taking HRT had lower incidence of cardiovascular disease

Q. Does HRT reduce risk of CVD?

A. NO – women on HRT tended to be from higher socioeconomic groups and have a healthier nutrition / exercise regime (i.e. coincidence effect)

Multilinear Regression

- The response of a dependent variable (y) is not always related to only one explanatory variable (x)



- Commonly, y will be influenced by a number of other variables

Multilinear Regression

- Multilinear regression accounts for several explanatory variables to predict the outcome of a dependent variable

Example

When determining the effect of FEV_1 decline on anxiety/depression scores in COPD, other factors that could influence anxiety/depression must be accounted for;

E.g. Age, sex, BMI, current lung function, symptoms, smoking status, socioeconomic status etc.

AGREEMENT

https://www.von.gov.ng/

Agreement

- Agreement refers to the **degree of concordance** between two (or more) sets of measurements

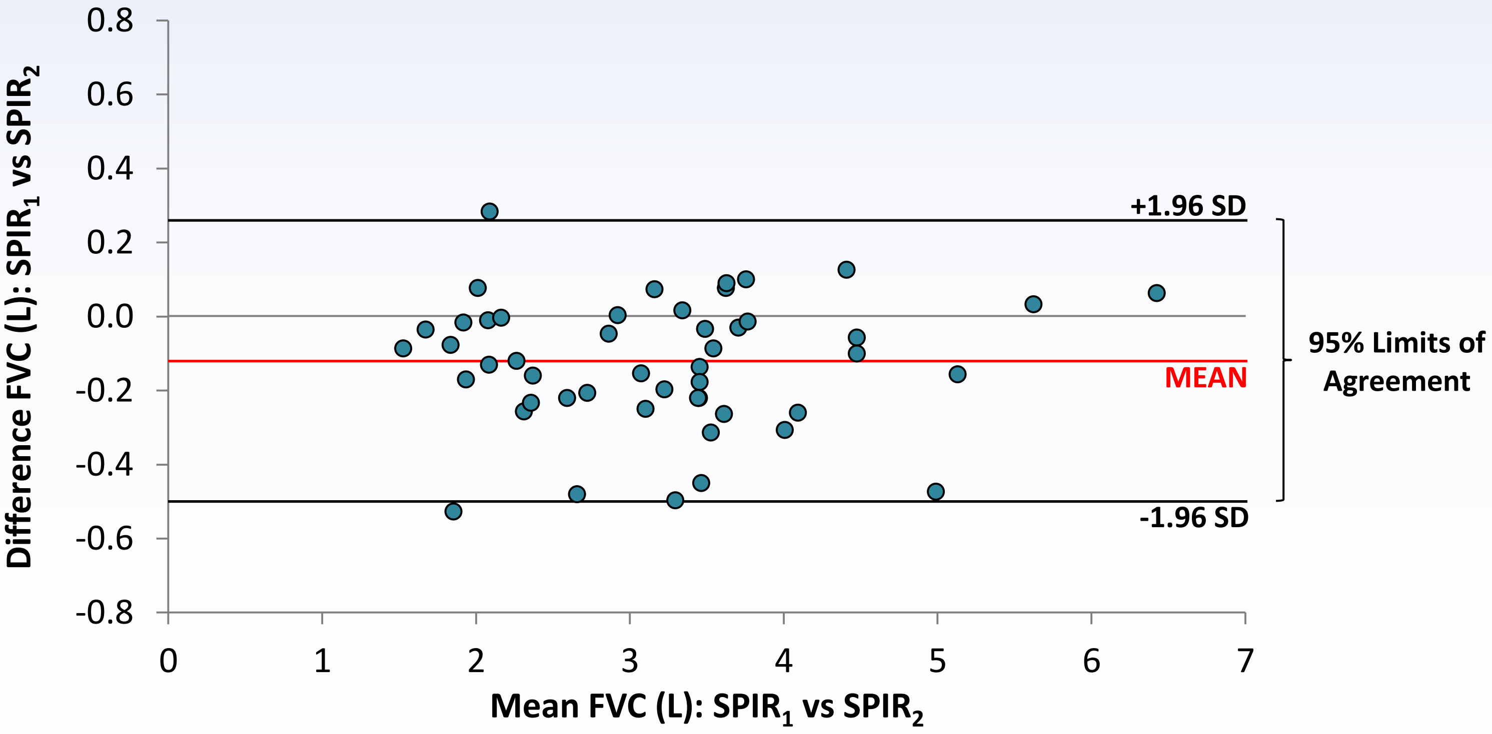
- Statistical methods to test agreement are used to;
  - Determine whether one technique for measuring a variable can substitute another
    - **e.g. multichannel as a substitute for polysomnography**
  - Assess inter-rater variability
    - **e.g. human vs auto-scoring for sleep diagnostics**

Bland-Altman Plots

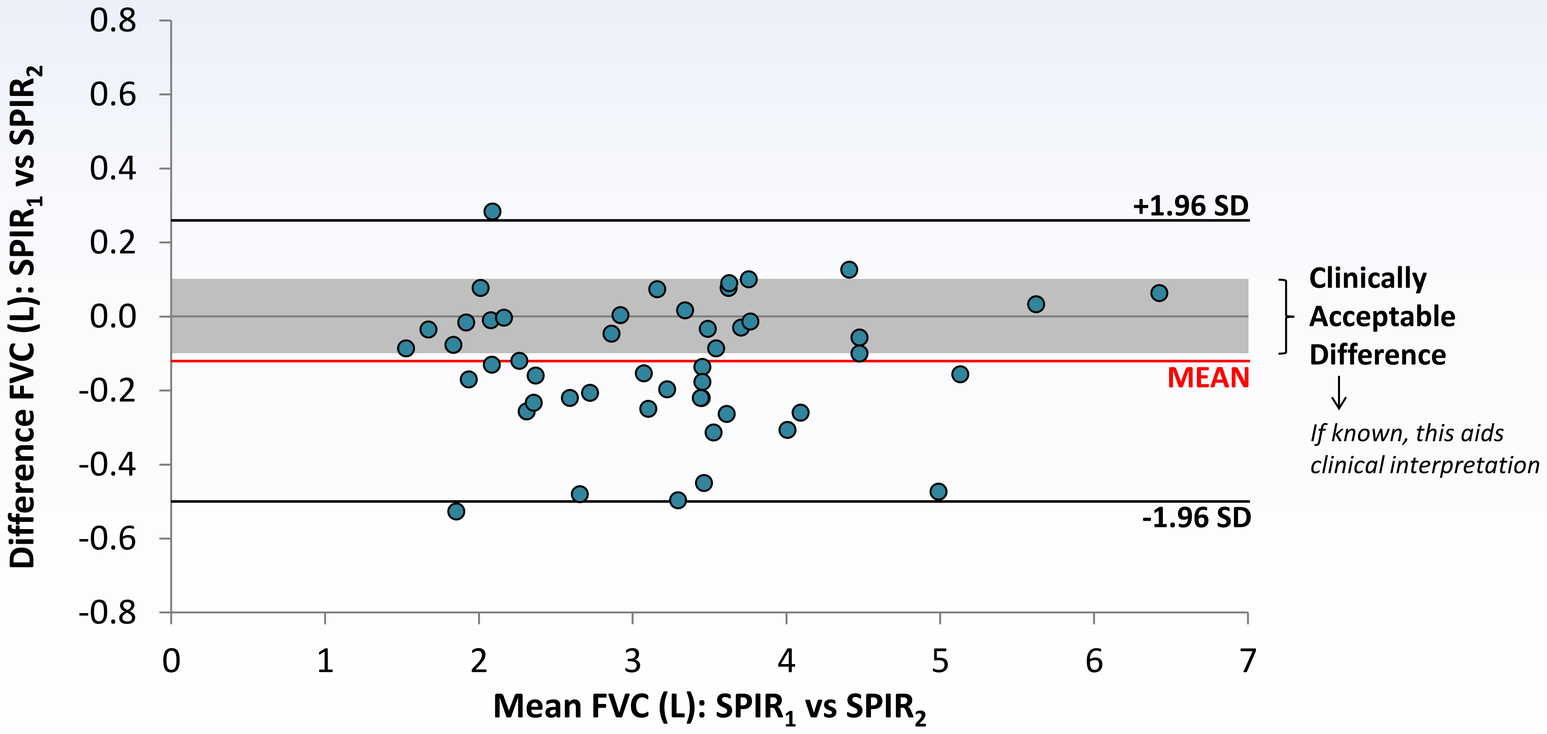
- Used to assess agreement between **two techniques** that measure the **same parameter**

- A scatter plot of the **mean** of the two measurements (x-axis) against the **difference** between the two measurements (y-axis)

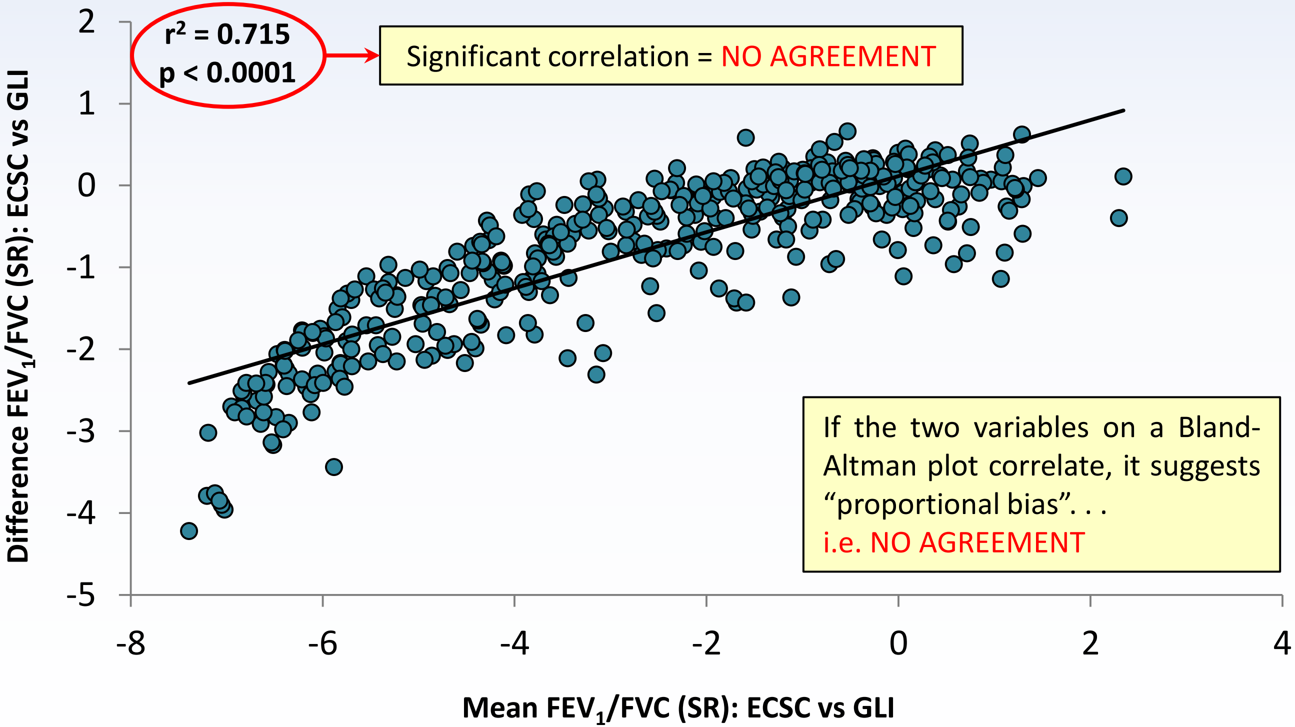
Bland-Altman Plots



Bland-Altman Plots



Proportional Bias



Intra-Class Correlation Coefficient

- Quantifies agreement of between $\geq$**2 observers** measuring the same **continuous** variable

- Calculations are complex →

  Different formulas for different applications

- Generates a number **0 – 1**

- Closer to 1, the stronger the agreement

$$r = \frac{1}{Ns^2} \sum_{n=1}^{N} (x_{n,1} - \bar{x})(x_{n,2} - \bar{x}),$$

where

$$\bar{x} = \frac{1}{2N} \sum_{n=1}^{N} (x_{n,1} + x_{n,2}),$$

$$s^2 = \frac{1}{2N} \left\{ \sum_{n=1}^{N} (x_{n,1} - \bar{x})^2 + \sum_{n=1}^{N} (x_{n,2} - \bar{x})^2 \right\}$$

https://en.wikipedia.org/wiki/Intraclass_correlation

Intra-Class Correlation Coefficient

ICC Value	Strength of Agreement
< 0.5	Poor
0.5 – 0.75	Fair
0.75 – 0.9	Good
> 0.9	Excellent

Koo, TK, Li MY. *J Chiropr Med* 2016; 15(2): 155-63

Cohen's Kappa Test

- Compares **binary nominal data** between **two observers**
  - E.g. Two consultants determining if patients should start treatment (**YES/NO**)

N = 50	YES *cons1*	NO *cons1*
YES *cons2*	20	10
NO *cons2*	5	15

$$K = \frac{P_o - P_e}{1 - P_e}$$

P_o = rate of *observed* YES/NO agreement
= (20 + 15) / 50 = **0.70**

Free Cohen's Kappa Calculator: https://idostatistics.com/cohen-kappa-free-calculator/

Cohen's Kappa Test

- Compares **binary nominal data** between **two observers**
  - E.g. Two consultants determining if patients should start treatment (**YES/NO**)

N = 50	YES *cons1*	NO *cons1*
YES *cons2*	**20**	**10**
NO *cons2*	**5**	**15**

$$K = \frac{P_o - P_e}{1 - P_e}$$

P_e = The probability of chance-*expected* agreement

Cons1 said Yes to 25/50 images, or 50% (0.5)
Cons2 said Yes to 30/50 images, or 60% (0.6)

The total probability of the consultants both saying **YES** randomly is 0.5 x 0.6 = **0.30**

Cons1 said No to 25/50 images, or 50% (0.5)
Cons2 said No to 20/50 images, or 40% (0.4)

The total probability of the consultants both saying **NO** randomly is 0.5 x 0.4 = **0.20**

P_e = 0.30 + 0.20 **= 0.50**

Free Cohen's Kappa Calculator: https://idostatistics.com/cohen-kappa-free-calculator/

Cohen's Kappa Test

- Compares **binary nominal data** between **two observers**
  - E.g. Two consultants determining if patients should start treatment (**YES/NO**)

N = 50	**YES** *cons1*	**NO** *cons1*
YES *cons2*	**20**	**10**
NO *cons2*	**5**	**15**

$$K = \frac{P_o - P_e}{1 - P_e} = \frac{0.7 - 0.5}{1 - 0.5} = \mathbf{0.4}$$

$\mathbf{P_o} = 0.70$

$\mathbf{P_e} = 0.50$

Free Cohen's Kappa Calculator: https://idostatistics.com/cohen-kappa-free-calculator/

Kappa Interpretation

κ Value	Strength of Agreement
0.01 – 0.20	Slight
0.20 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost Perfect
1.00	Perfect

Landis JR & Koch GG. *Biometrics* 1977; 33: 159-174

Kappa Variants

- Other Kappa tests are available for depending on the number of observers and type of data (nominal vs ordinal);

Type of Variable	Number of Observers	Test
Nominal	2	Cohen's Kappa
e.g. Disease present? YES/NO	>2	Fleiss' Kappa
Ordinal	2	Weighted Kappa
e.g. Disease severity	>2	Fleiss' Kappa

GROUP COMPARISON

https://www.von.gov.ng/

Group Comparison

- Methods for **comparing averages** between two or more groups of numerical or ordinal data

Need to know;

1. **Distribution** (only use parametric tests if **ALL** groups are normally distributed)

2. **Number** of groups (2 or >2)

3. Are the groups different subjects (**independent**) or the same subjects at different time points (**paired**)?

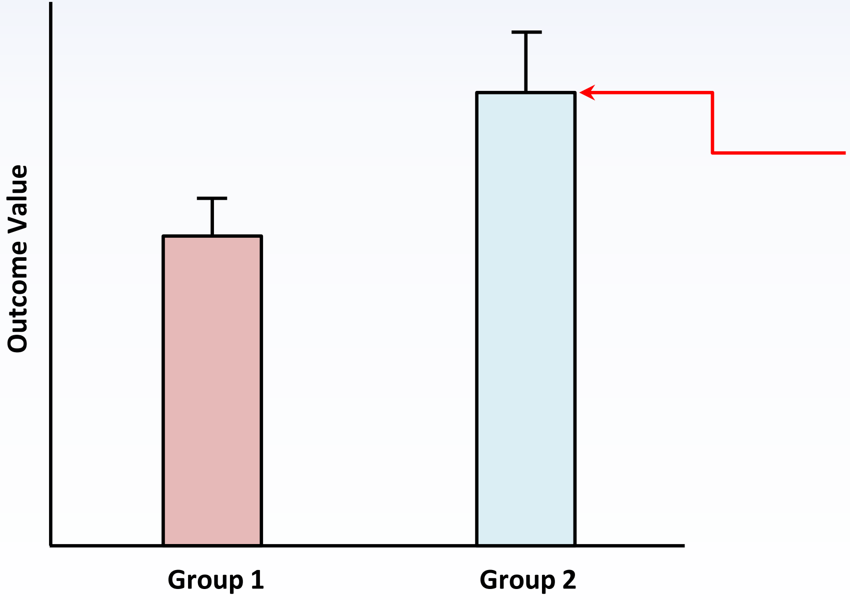
4. Is the result predictable (**p** vs **2xp**)?

Statistical Tests

Distribution	Number of Groups	Type of Group	Test
Normal **(ALL groups)**	2	Independent	t-Test
	2	Paired	Paired t-Test
	>2	Independent	ANOVA
	>2	Paired	Repeated Measure ANOVA
Not Normal **(ANY group)**	2	Independent	Mann Whitney-U
	2	Paired	Wilcoxon Signed-Rank
	>2	Independent	Kruskal-Wallis
	>2	Paired	Friedman’s
Ordered Alternatives **(e.g. COPD severities)**	>2	Independent	Jonckheere-Terpstra

Graphical Representation

- For normally distributed data, a histogram is conventional;

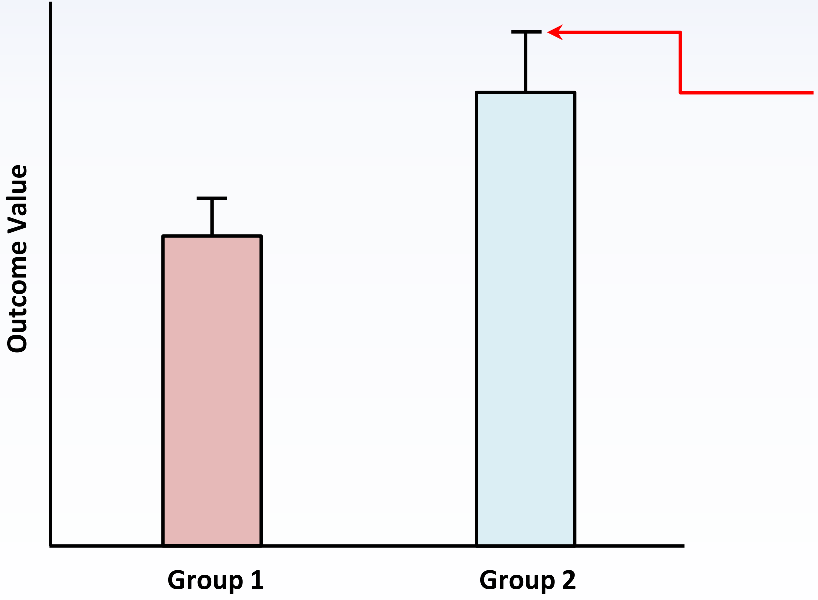


The top of the histogram is the **mean** of each group

$$MEAN = \frac{\textit{Sum of all values}}{\textit{Number of values}}$$

Graphical Representation

- For normally distributed data, a histogram is conventional;

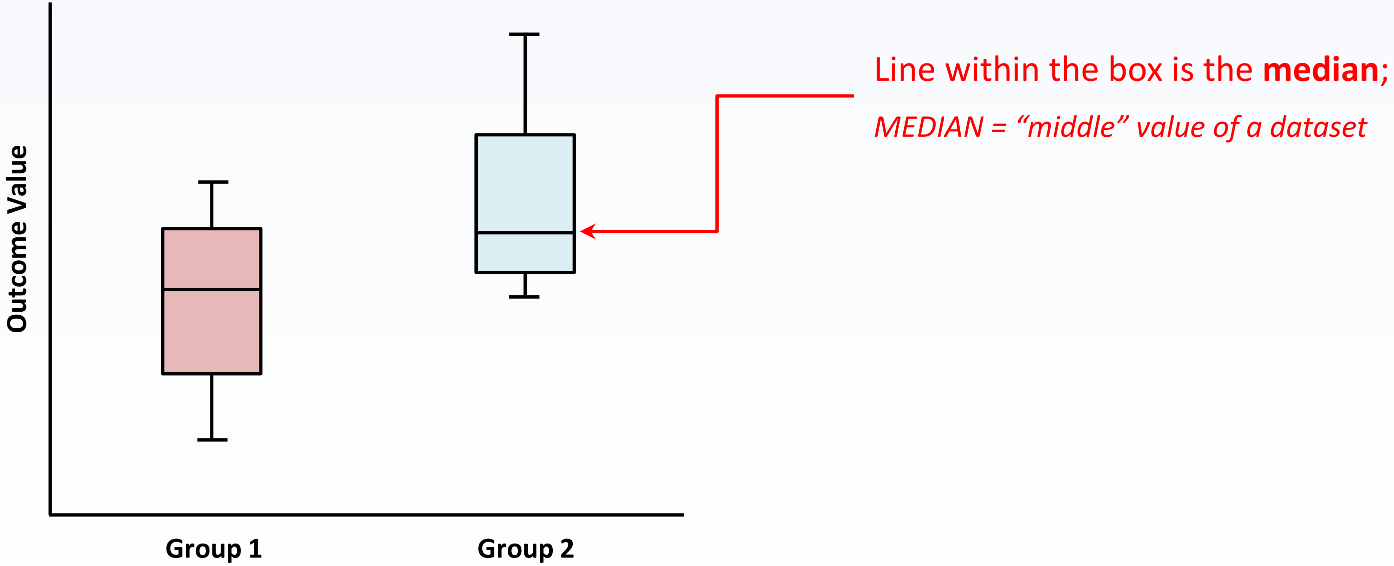


Error bars may either be;

Standard Deviation *– "variability" of data*
or
Standard Error *– accuracy of data (sample mean vs population mean)*

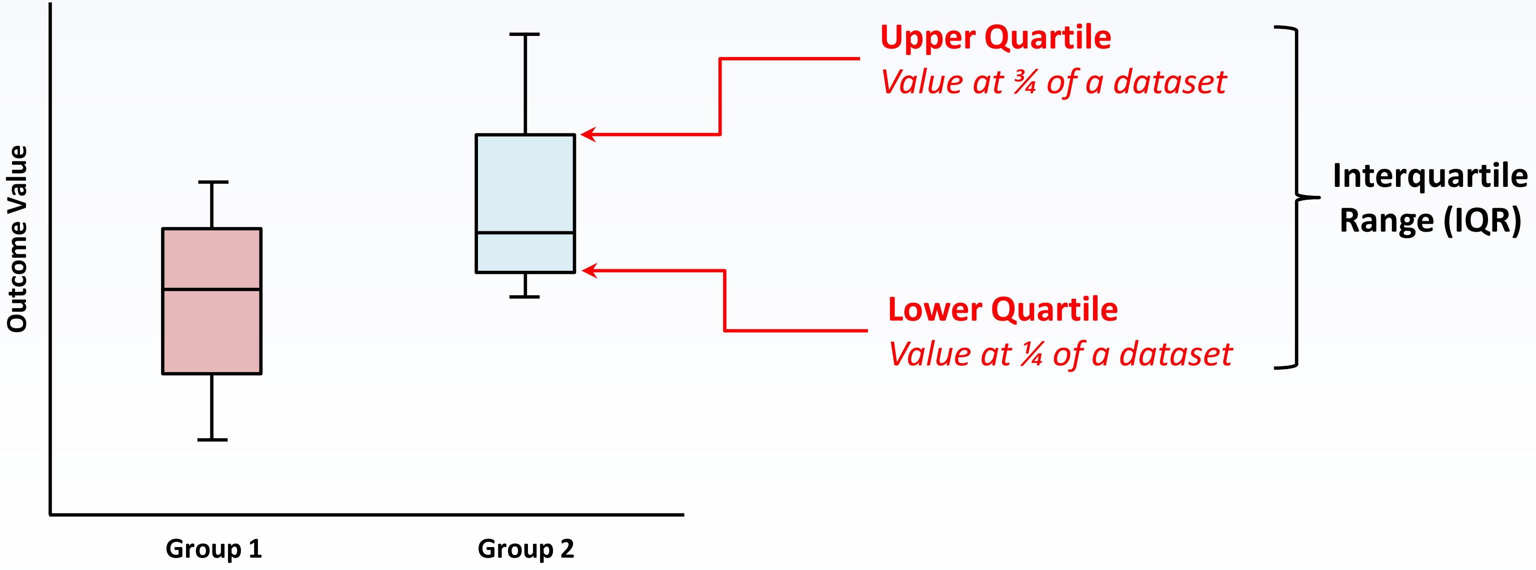
Graphical Representation

- For data that are not normally distributed, Box & Whisker plots are used;



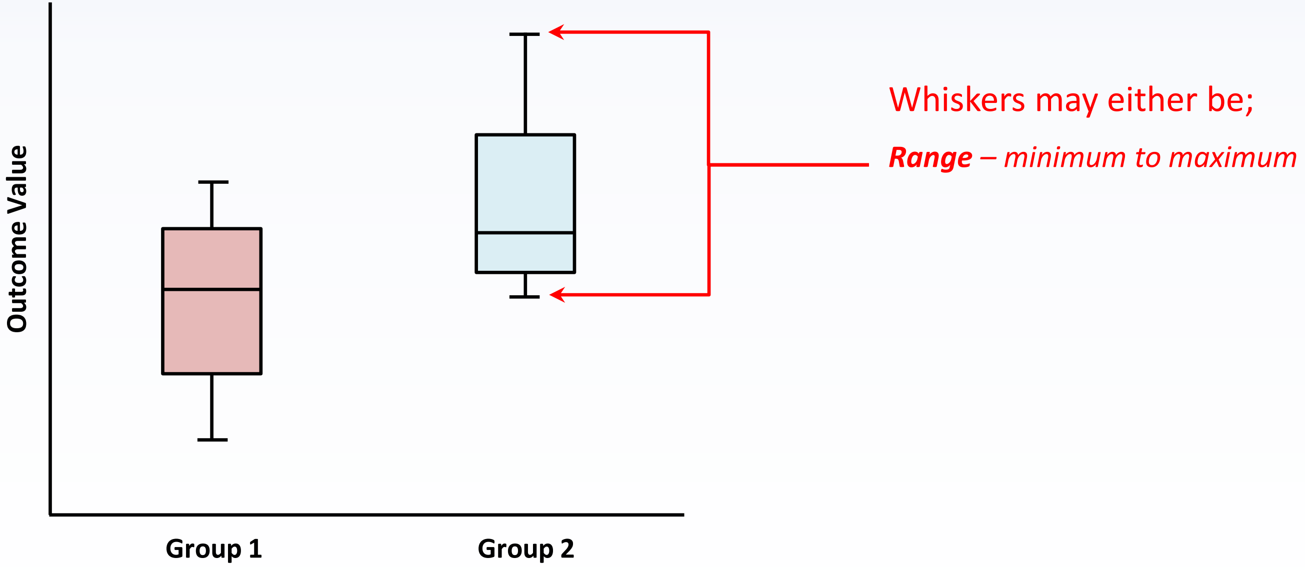
Graphical Representation

- For data that are not normally distributed, Box & Whisker plots are used;



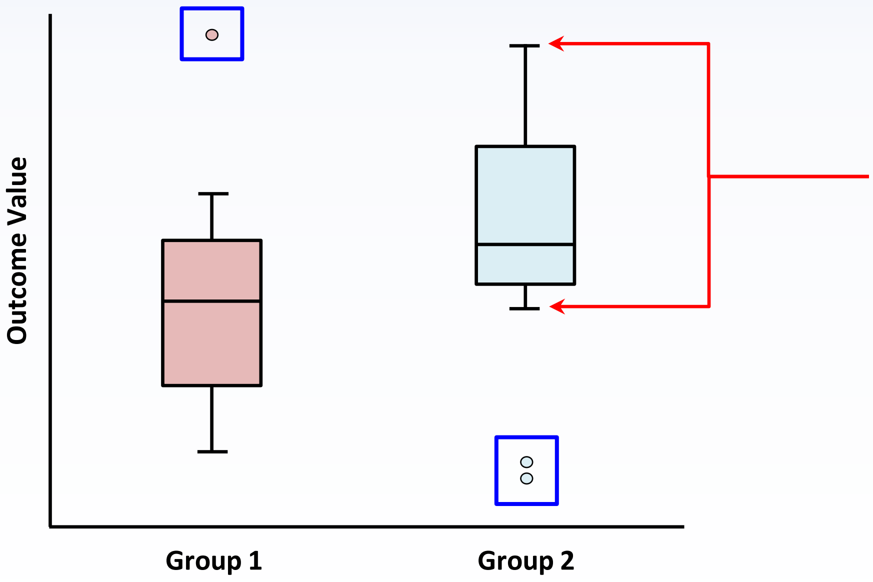
Graphical Representation

- For data that are not normally distributed, Box & Whisker plots are used;



Graphical Representation

- For data that are not normally distributed, Box & Whisker plots are used;



Whiskers may either be;

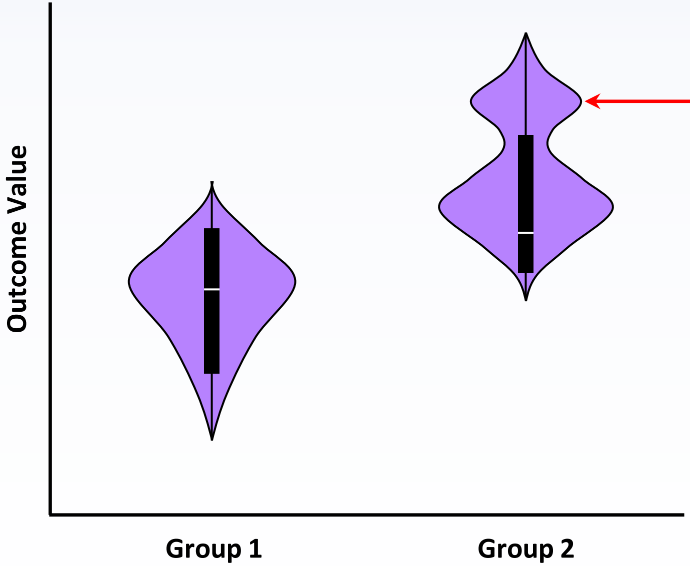
Range *– minimum to maximum*

or

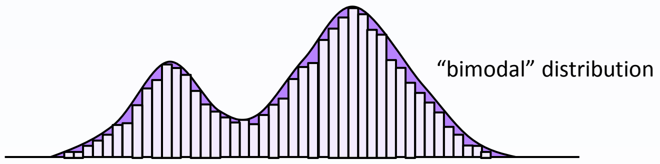
1.5 x IQR *– useful to highlight* ***outliers***

Graphical Representation

- A **Violin Plot** is created by superimposing the **distribution** over the box & whiskers;



“Violin” is the distribution curve (from histogram) rotated 90º;



Useful in visually demonstrating that some values occur more frequently (e.g. CPAP compliance)

Comparing Nominal Data

- For nominal data (e.g. sex), the following tests are recommended;

	Group 1	Group 2
Males	52	35
Females	24	60

Fisher's Exact test
(more accurate when total sample < 1,000*)

Chi-Squared test of independence
(more easily applied to total samples ≥ 1,000*)

Both can be expanded beyond a 2x2 table but not all software allows this for the Fisher's test

* http://www.biostathandbook.com/small.html

Comparing *Paired* Nominal Data

- For example, presence of disease or symptoms (YES/NO) before and after treatment in the same patients;

	After: Present	**After:** Absent
Before: Present	**80**	**16**
Before: Absent	**6**	**59**

McNemar's Test

E.g. Patients with airflow obstruction (defined by FEV_1/FVC LLN) before and after salbutamol 2.5mg nebuliser

TEST CHARACTERISTICS

https://www.von.gov.ng/

Test Characteristics

	Disease present (+)	Disease absent (–)	Totals
Test result positive (+)	**TP**	**FP**	TP + FP
Test result negative (–)	**FN**	**TN**	FN + TN
Totals	TP + FN	FP + TN	

- Important method for assessing the **accuracy of a diagnostic test**
- The diagnostic test under investigation is called the **index test**

Test Characteristics

	Disease present (+)	Disease absent (−)	Totals
Test result positive (+)	**TP**	**FP**	TP + FP
Test result negative (−)	**FN**	**TN**	FN + TN
Totals	TP + FN	FP + TN	

- Important method for assessing the **accuracy of a diagnostic test**
- The diagnostic test under investigation is called the **index test**
- This is compared to the **reference standard** (usually the best test currently available)

Test Characteristics

	Disease present (+)	Disease absent (−)	Totals
Test result positive (+)	**TP**	**FP**	TP + FP
Test result negative (−)	**FN**	**TN**	FN + TN
Totals	TP + FN	FP + TN	

TP = True Positive

Index test correctly identifies disease

FP = False Positive

Index test incorrectly identifies disease

TN = True Negative

Index test correctly identifies no disease

FN = False Negative

Index test incorrectly identifies no disease

Test Characteristics

	Disease present (+)	Disease absent (−)	Totals
Test result positive (+)	**TP**	**FP**	TP + FP
Test result negative (−)	**FN**	**TN**	FN + TN
Totals	TP + FN	FP + TN	

Sensitivity

– True positive rate

% Patients with disease correctly identified by index test

– **TP/(TP+FN)**

Test Characteristics

	Disease present (+)	Disease absent (−)	Totals
Test result positive (+)	TP	**FP**	TP + FP
Test result negative (−)	FN	**TN**	FN + TN
Totals	TP + FN	FP + TN	

Specificity

- True negative rate

% Patients with no disease correctly identified by index test

- **TN/(FP+TN)**

Test Characteristics

	Disease present (+)	Disease absent (−)	Totals
Test result positive (+)	**TP**	**FP**	TP + FP
Test result negative (−)	**FN**	**TN**	FN + TN
Totals	TP + FN	FP + TN	

Positive Predictive Value

- Proportion of people with a positive test result who actually have the disease
- **TP/(TP+FP)**

Test Characteristics

	Disease present (+)	Disease absent (−)	Totals
Test result positive (+)	TP	FP	TP + FP
Test result negative (−)	**FN**	**TN**	FN + TN
Totals	TP + FN	FP + TN	

Negative Predictive Value

- Proportion of people with a negative test result who do not have the disease
- **TN/(FN+TN)**

Test Characteristics

- Sensitivity and specificity are **fixed** for a particular test
- PPV and NPV for a particular type of test depend upon the **prevalence** of a disease in a population
- Population could be the general public (**screening tool**) or a patient group (**diagnostic tool**)

Clinical Practice

- PPV and NPV are often more useful in practice than sensitivity/ specificity
  - *If a disease is very rare, sensitivity/specificity can be high but PPV can still be low*

Example

Current screening tests for HIV have **high sensitivity** and **high specificity**. However, the low prevalence of HIV in the general population cannot justify universal screening since the majority of positive tests would actually be false positives (i.e. FP > TP = **low PPV**)

Clinical Practice

Challenge Testing

- **Methacholine** - high sensitivity, lower specificity → **High NPV**
  - Better at *excluding* asthma
  - Confident a negative result **is not** asthma
- **Mannitol** - high specificity, lower sensitivity → **High PPV**
  - Better at *confirming* asthma
  - Confident a positive result **is** asthma

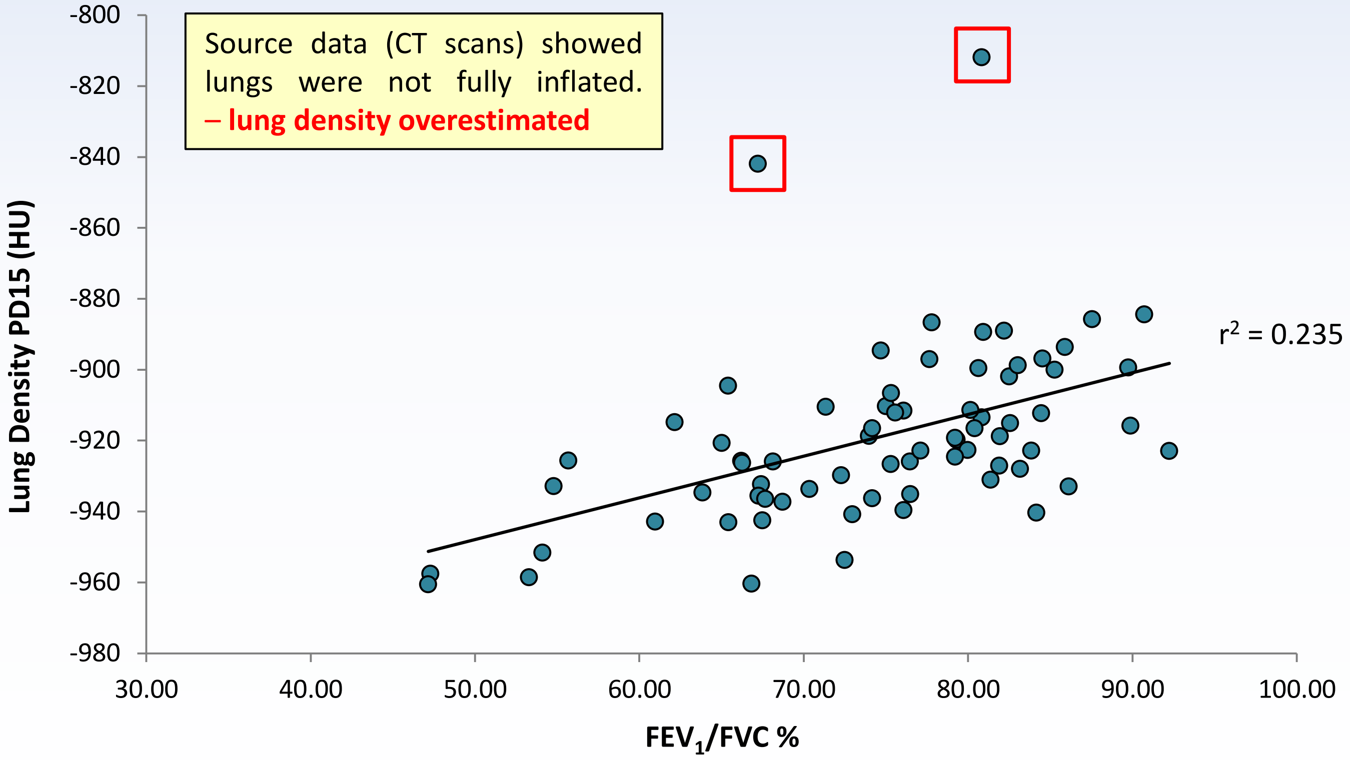
CONSIDERATIONS

https://www.von.gov.ng/

Data Validation

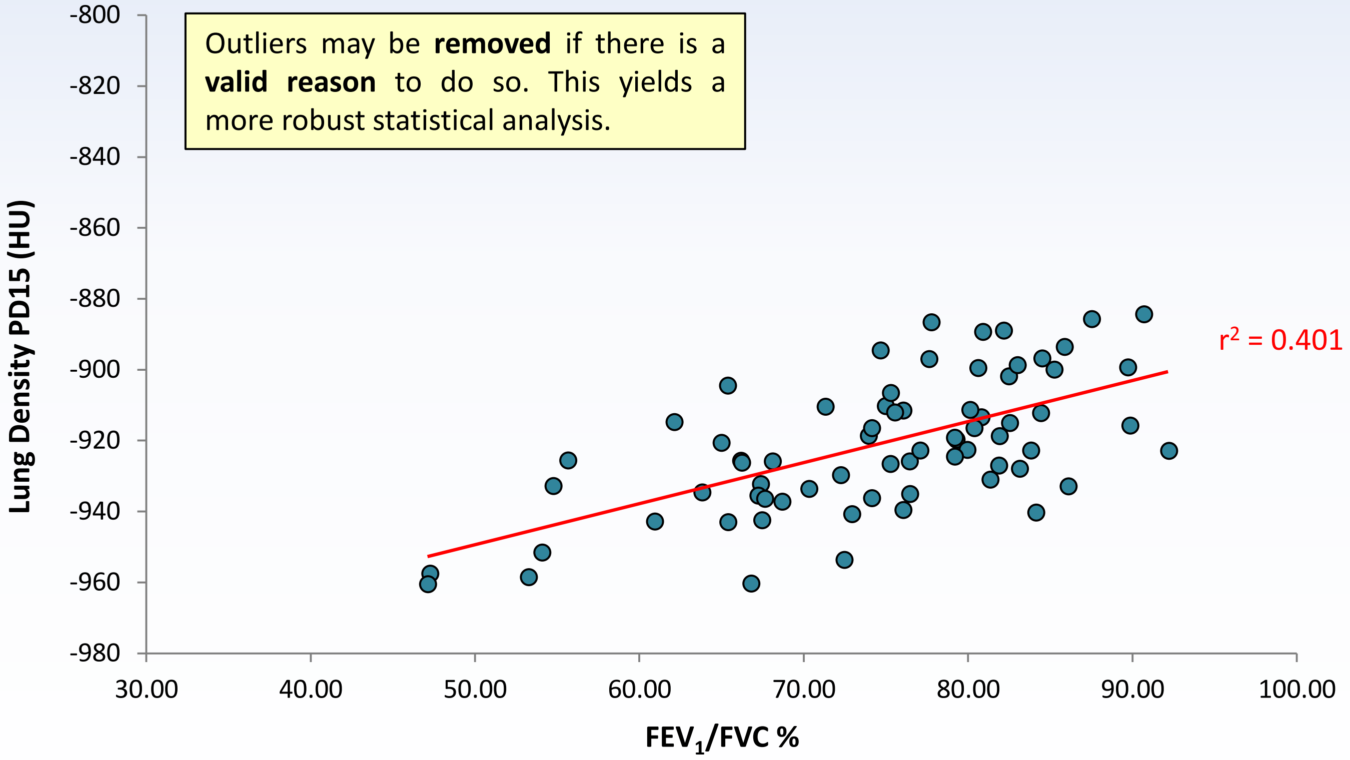
- **Transcription errors** may occur when copying data from the source to a database
- Periodic **data verification** by the investigator or an independent party helps ensure data accuracy
- Often a small sample of the data (e.g. 10%) is validated

Outliers


Source data (CT scans) showed lungs were not fully inflated.
– lung density overestimated
Lung Density PD15 (HU)
-800
-820
-840
-860
-880
-900
-920
-940
-960
-980
30.00
40.00
50.00
60.00
70.00
80.00
90.00
100.00
FEV1/FVC %
r2 = 0.235

Outliers

Outliers may be **removed** if there is a **valid reason** to do so. This yields a more robust statistical analysis.



Unexpected Outcomes

- An effect can be observed that is not associated with the original question
- If the effect may be important, it is good scientific practice to repeat the experiment with a new hypothesis
- It is **bad scientific practice** to modify a hypothesis after statistical analyses or “fish” for data if H_0 is true

Interpretation

- The process of “making sense” of the data
- Do so **accurately** and **impartially**
  - *Do not fudge data or use inappropriate statistics to reach p < 0.05*

“p=hacking”

- **Statistical vs clinical** significance
  - *How might your findings **really** influence patient care?*
- Do not overstate conclusions

SCIENCE IS FUNDAMENTALLY A MORAL ENTERPRISE, FOLLOWING THE MORAL IMPERATIVE TO SEEK THE TRUTH

– George Lakoff –

Summary

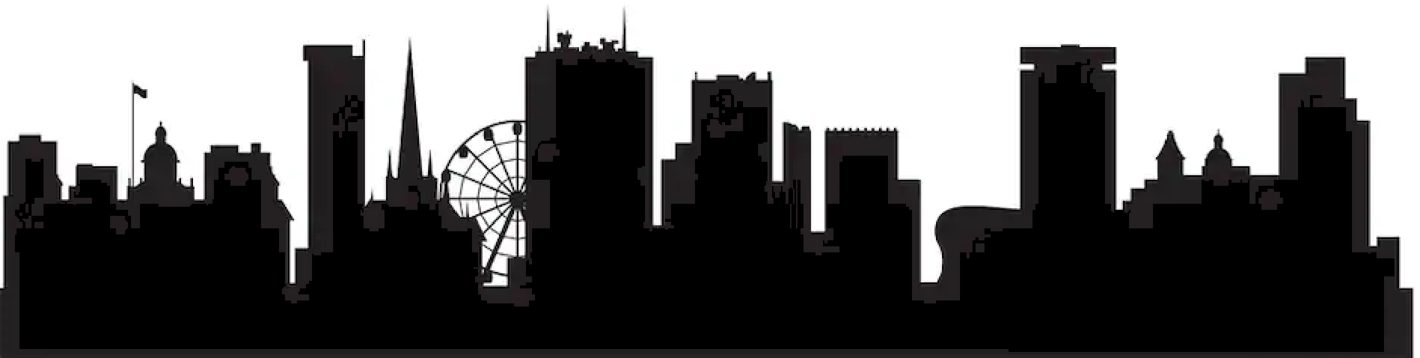
- Start with a question → background research → hypothesis
- Formulate the methodology and statistical tests around this
- Understand the type of data
- Gather data carefully and validate periodically
- Interpret impartially and in relation to clinical meaning
- **Seek advice from a statistician!**

RESEARCH & INNOVATION COMMITTEE

@ARTP_Research   research@artp.org.uk

Please contact us. We happy to provide guidance and support in all aspects of respiratory and sleep research including;

- **Research design / methodology**
- **Statistical analyses**
- **Ethics applications**
- **Funding streams**
- **Abstract preparation**
- **Conference presentations**



ARTP