Statistics

Kristof Reid

Assistant Professor

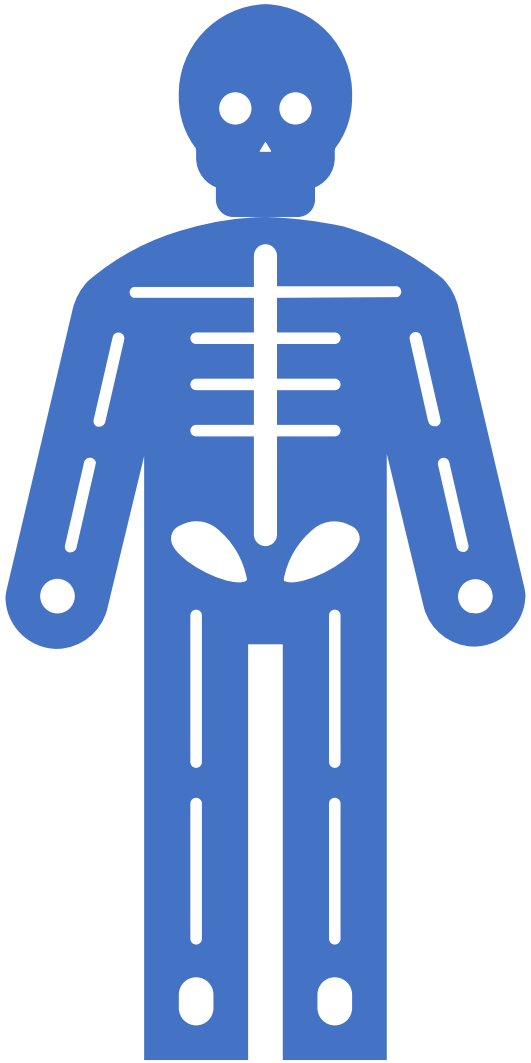
Medical University of South Carolina

Financial Disclosures

- None

Further Disclosures

- I am not a statistician
- I do like making clinical decisions based on appropriately interpreted data



Learning Objectives

- Understand why knowing statistics is important
- Understand the basic principles and statistical test
- Understand common statistical errors in the medical literature

Why should I care?

Indications for statistics

Topics in Training

A National Survey of Orthopaedic Residents Identifies Deficiencies in the Understanding of Medical Statistics

Ibukunoluwa Araoye, MD, Jun Kit He, BS, Scott Gilchrist, MD, Trevor Stubbs, MD, Gerald McGwin Jr., MS, PhD, and Brent A. Ponce, MD, on behalf of the Collaborative Orthopaedic Educational Research Group*

47% unable to determine study design

31% did not understand p values

38% did not understand sensitivity and specificity

83% could not use odds ratios

Araoye et al, JBJS 2020;102(5):e19

Statistics Corner

Misuse of statistics in surgical literature

Matthew S. Thiese[1], Brenden Ronna[1], Riann B. Robbins[2]

[1]Rocky Mountain Center for Occupational & Environment Health, Department of Family and Preventive Medicine, [2]Department of Surgery, School of Medicine, University of Utah, Salt Lake City, Utah, USA

Correspondence to: Matthew S. Thiese, PhD, MSPH. Rocky Mountain Center for Occupational & Environment Health, Department of Family and Preventive Medicine, School of Medicine, University of Utah, 391 Chipeta Way, Suite C, Salt Lake City, UT 84108, USA. Email: matt.thiese@hsc.utah.edu.

50% (or more!) of clinical research publications have at least one statistical error

Thiese et al, Journal of thoracic disease, 2016-08, Vol.8 (8), p.E726-E730

■ INSTRUCTIONAL REVIEW: GENERAL ORTHOPAEDICS

A systematic survey of the quality of research reporting in general orthopaedic journals

17% of conclusions not justified by results

39% of studies used the wrong analysis

Parsons et al, J Bone Joint Surg Br. 2011;93-B(9):1154-1159.

Are these two columns different?

You have been asked by your insurance carrier to prove that your total hip patient outcomes are not statistically different than your competitor next door.
Column A is the patient reported score for you and column B for your competitor

How would you prove you are different or better?

Column A	Column B
19	10
12	11
10	9
20	19
10	17
14	10
21	19
24	13
12	18
17	8
12	17
17	10
9	9
24	10
21	9
18	18
13	4
18	8
12	14
15	15
14	17

Are they still the same?

Column A: Mean=15

Column B: Mean=12

SD=4

We want to know if we're making a difference

What is statistics?

a. A class in medical school that I slept through during first year

b. Something for epidemiology nerds

c. A mathematical discipline that gives us tools to make decisions based on incomplete samples.

Answer = c

Who needs to know statistics?

a. Statisticians

b. Journal Editors

c. Anyone intending to change their practice based on the medical literature

d. Anyone interested in interpreting Patient Satisfaction Scores and Surveys

Answer = All of the above

Why do we need to know statistics?

- Our intuition is often wrong
- Potential errors in the medical literature
- Understanding what statistical significance really means

Answer = Probably all the above

Intuition about statistics

- “Trending to significance”
- "The P-Value is low, but it’s a small sample size”
- Statistics exists because we can't do the analysis without tools

BMJ

BMJ 2014;348:g2215 doi: 10.1136/bmj.g2215 (Published 31 March 2014)

RESEARCH METHODS & REPORTING

Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data

- "trend to significance" is shaky at best

Wood, F. (2014). BMJ : British Medical Journal, 348(mar31 2), g2215–g2215

But, this is just for the OITE, right?

- Most surgeons now do procedures they never learned in residency.
- How do you decide what to add to your practice?
- How can you judge results for a procedure you do rarely?

What do we need to know?

- Interpret correctly reported statistics.
  - Descriptive statistics
  - Test results
- Have a feel for common errors

Statistics Background

Statistical Methods

- Underlying assumptions
- Common Tests
- Uncommon Tests
- Study Design

Variable: quantity that can take various values

- Types of Variables
  - Binary – only 2 possible categories e.g. yes or no, left or right
  - Categorical - individual belongs to one of a number of distinct categories e.g. Blood types
  - Numerical - values are discrete or continuous e.g. weight
  - Ordinal – categories placed in distinct order e.g. Patient Satisfaction (high, neutral, low)
  - Continuous – counts or measures that can have an infinity of values e.g. T score

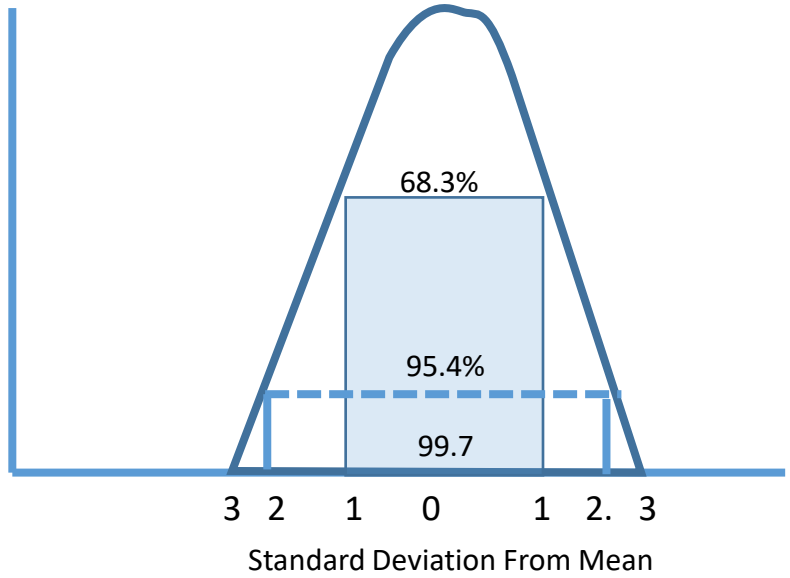
Variable: quantity that can take various values

- Categorical
  - Ordinal (ordered) e.g. disease stages 1, 2or 3
  - Nominal (unordered) e.g. blood groups A, B, Ab and O

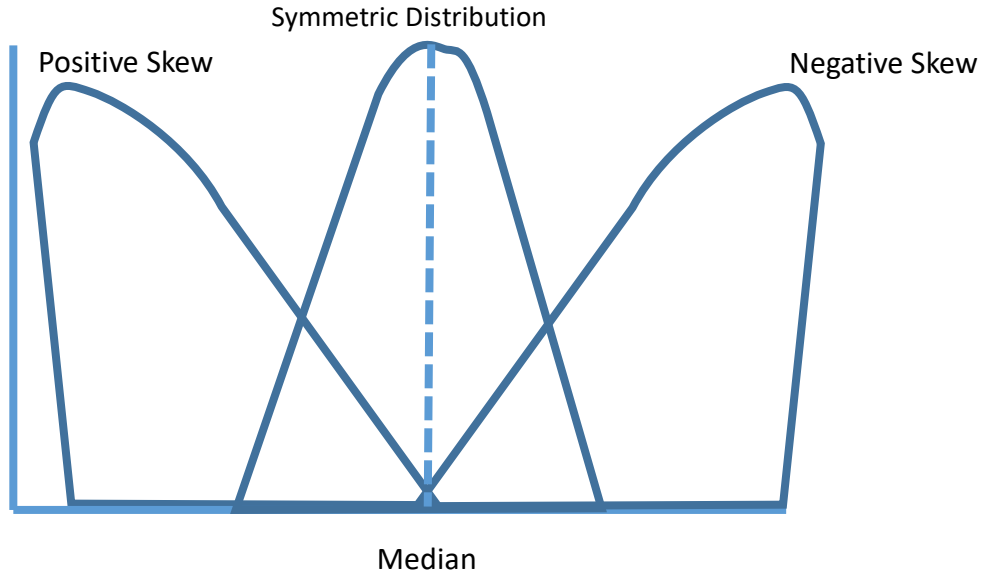
- Numerical
  - Discrete (integer) e.g. visit number
  - Continuous e.g. blood pressure in mmHg

Data distribution

Normal = data distributed about the mean is symmetrical



Skewed data distortion or asymmetry in a symmetrical bell curve, or normal distribution set of **data**.



The Null Hypothesis (H_o)

- Formalized expression of the research idea containing both quantitative and qualitative components
- Must be clear, unequivocal and answerable
- The basic assumption for statistical testing
  - There are no differences between two measured groups – two measured groups are equally effective
  - Two measured groups come from the same population

Summary Measures for Variables Types

- Categorical done on number per category e.g. percentage
- Numerical is done with
  - Arithmetic mean = adding all the observations and dividing this sum by the number in the dataset, best for symmetrical distribution of data
  - Median = observation which falls in the middle of the set of observations when they are arranged in increasing order of magnitude, best for skewed data as less influenced by outliers
  - Mode = most commonly observed value. Equals the mean and median in a normal distribution. Helpful when interpreting in bar graphs or other noncontinuous data

- Standard deviation (SD) is the average of the deviations of all the observations from the mean of the observations

Summary Measures for Variables Types

Odds Ratio

A measure of how strongly an event is associated with exposure.

- it is a ratio of the odds of the event occurring in an exposed group versus the odds of the event occurring in a non-exposed group.

- commonly are used to report case-control studies as it helps identify how likely an exposure is to lead to a specific event.

- The larger the odds ratio, the higher odds that the event will occur with exposure.

- Odds ratios smaller than one imply the event has fewer odds of happening with the exposure.

Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry. 2010 Aug;19(3):227-9

Summary Measures for Variables Types

Odds Ratio

- Relative probability
- Comes with a confidence interval
  - If the confidence interval for the odds ratio includes the number 1 then the calculated odds ratio would not be considered statistically significant.

Summary Measures for Variables Types

Confidence Interval

The confidence interval indicates the level of uncertainty around the measure of an effect (precision of the effect estimate)

The **confidence** level is the probability that the **confidence interval** contains the true **odds ratio**. If the study was repeated and the range calculated each time, you would expect the true value to lie within these ranges on 95% of occasions.

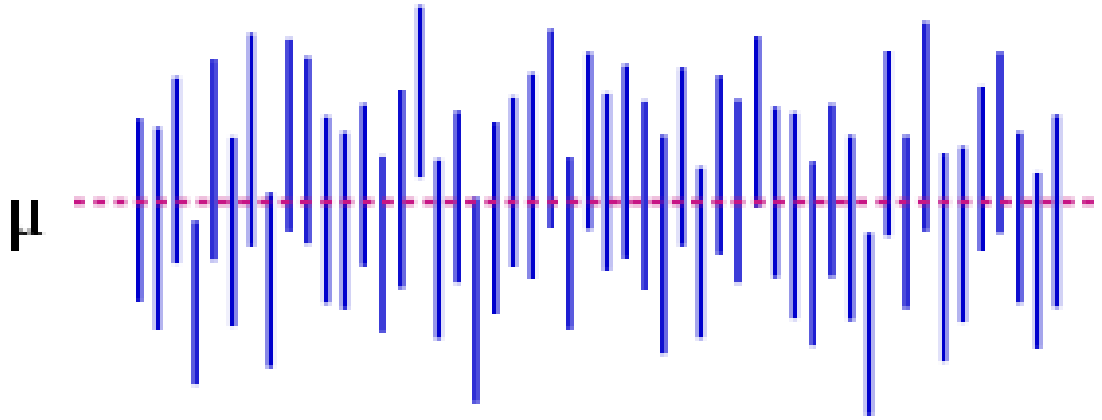
- the narrower the better

Confidence Intervals

The chance that the true value lies within a range of numbers.

Ho et Al, Nat Methods. 2019 Jul;16(7):565-566

Confidence Intervals

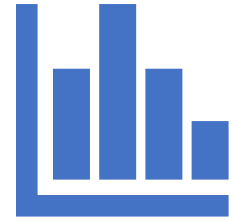


The blue vertical line segments represent 50 realizations of a confidence interval for the population mean μ, represented as a red horizontal dashed line; note that some confidence intervals do not contain the population mean, as expected.

Statistical Tests –The Launching Point

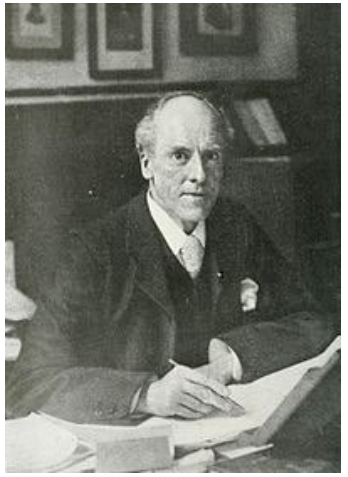
- Distinguish between results compatible with chance and those that no longer can be explained by chance

- Used to determine if the null hypothesis holds true



T-Test (Student T-Test)

- Published in English literature in 1908 by Gosset and Pearson
- Used to monitor the quality of Guinness Stout
- Used the the pseudonym Student
- Determines whether the **means** of two populations differ
- Assumes normal(ish) distribution

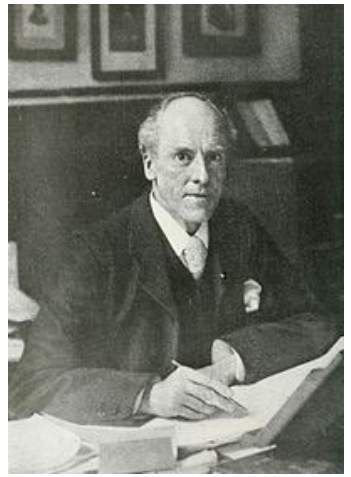


Karl Pearson

William Sealy Gosset

Chi square Test

- Published in 1900 by Pearson
- Determines whether two **categorical** variables differ
- Assumes normal(ish) distribution



Karl Pearson

Analysis of Variance (ANOVA)

Ronald Fisher

- Generalized t-test to test variance among and between **groups (2 or more)** or population means
- One way test used to compare means of two or more samples
  - Normal distribution, ordinal
- Two way test used to analyze two different categorical independent variables on one continuous dependent variable scale

Mann-Whitney U-Test

- Also known as Mann-Whitney-Wilcoxon Test, Wilcoxon rank sum test
- Non parametric - Not normally distributed
- Independent observations from each group
- Ordinal responses
- Small samples

Other Tests

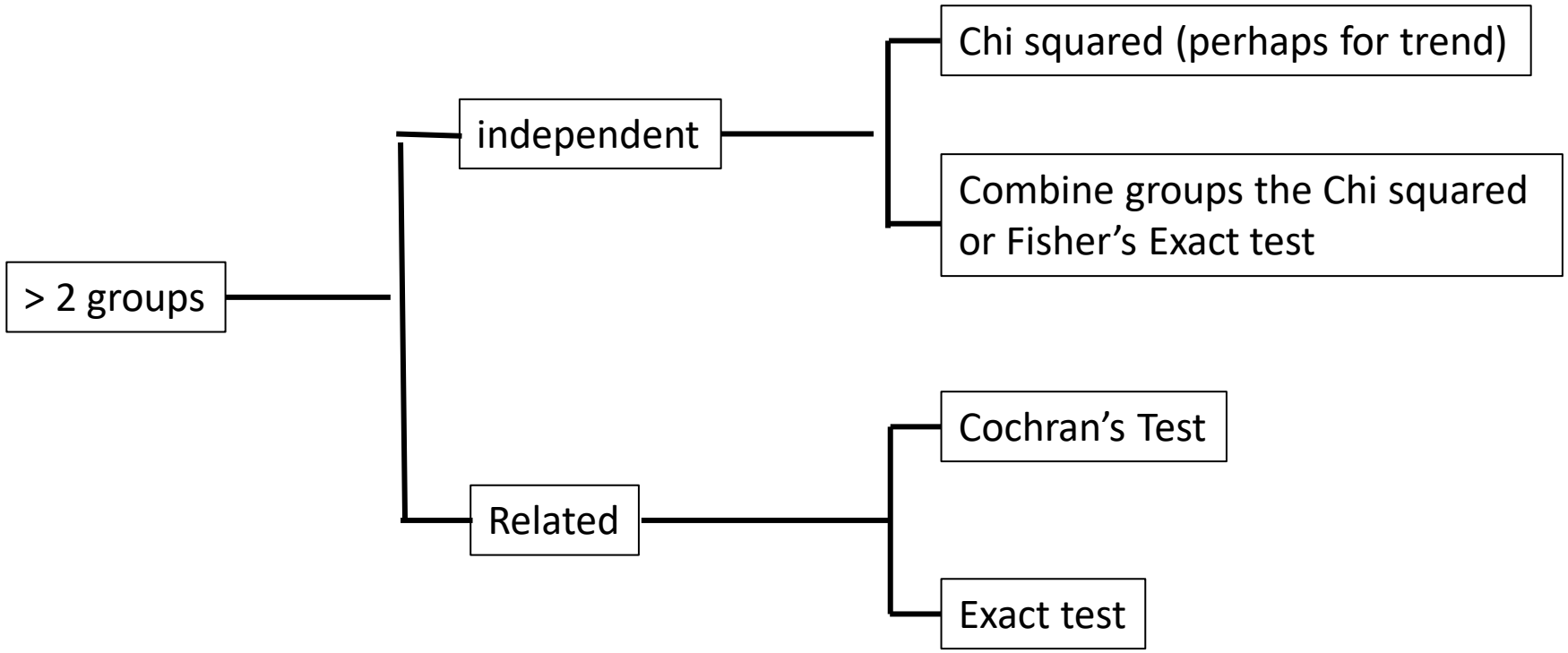
- McNemar's test – for paired nominal data
- Cochran's Q test – for non-parametric, matched sets of three or more frequencies or proportions
- Kruskal Wallis test – non-parametric for 2 or more independent samples of equal or different sizes
- Bonferroni's correction – correction for multiple comparison of data or hypotheses, lessen the chance of a Type 1 error

Choice of a Statistical Test to prove the Hypothesis

4 Questions to answer:

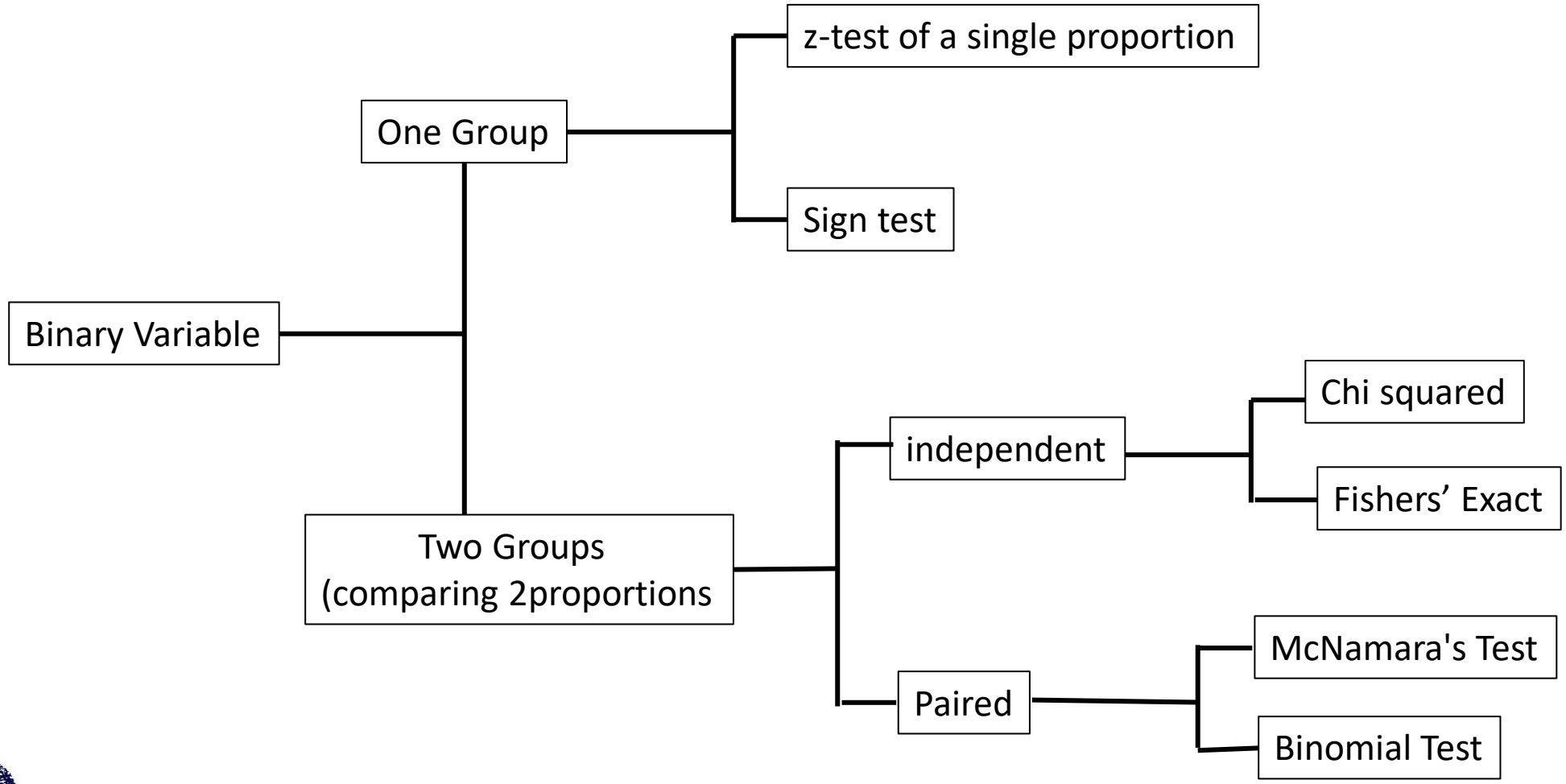
1. Categorical or numerical variables?
2. How many groups are compared?
3. Independent or related groups?
4. Are all assumptions underlying the proposed test satisfied?
    1. Test with no assumptions about distribution are non parametric
    2. Tests assuming normal distribution are parametric

Choice of Hypothesis Test

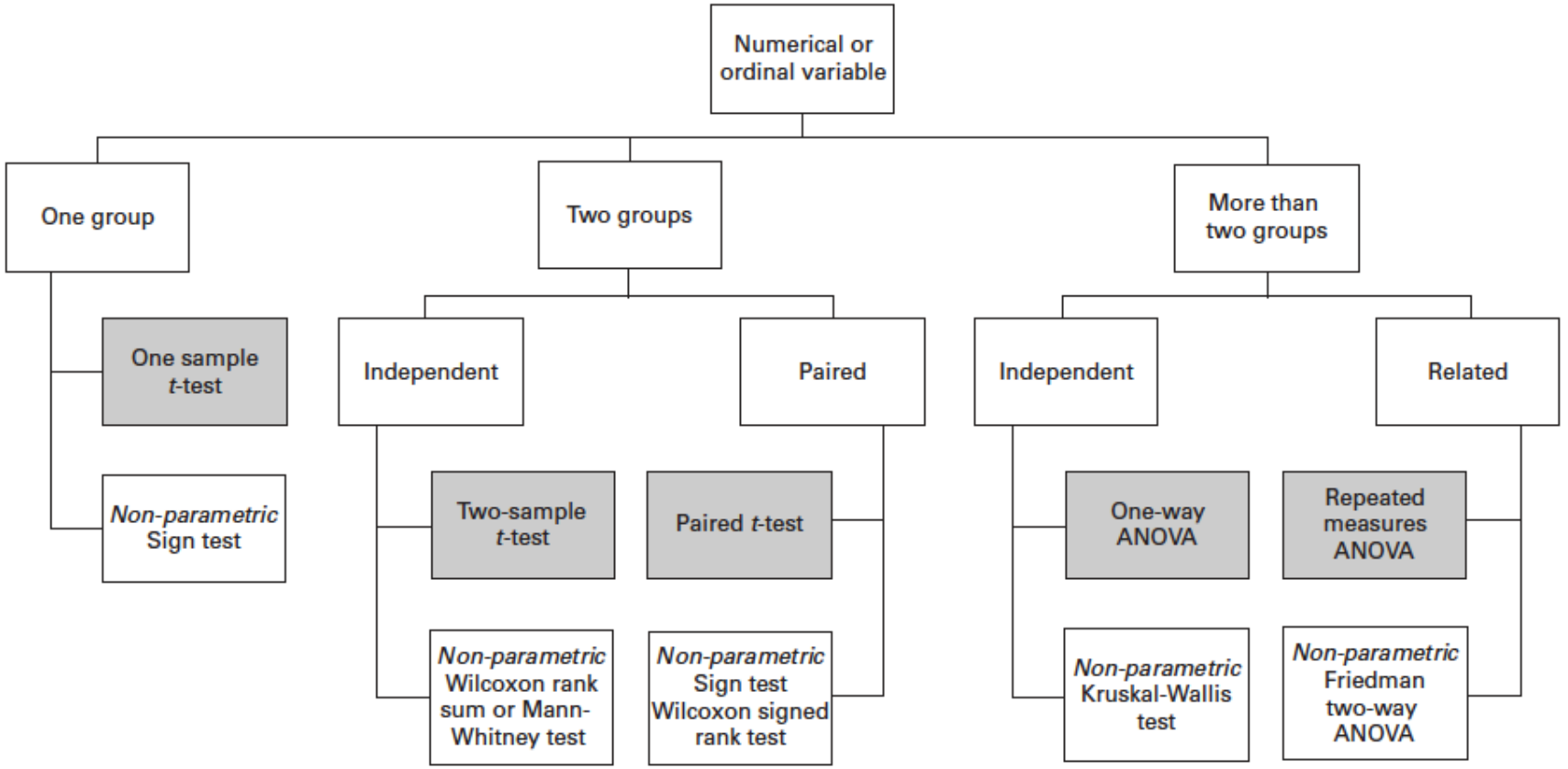


A. Petrie Statistics in orthopaedic papers VOL. 88-B, No. 9, SEPTEMBER 2006 1121

Choice of Hypothesis Test



Choice of Hypothesis Test



Flowchart indicating choice of test when the data are numerical (tests in shaded boxes require relevant assumptions to be satisfied) (ANOVA, analysis of variance).

A. Petrie Statistics in orthopaedic papers VOL. 88-B, No. 9, SEPTEMBER 2006 1121

P-Value

"The value for which P = 0.05, or 1 in 20, is 1.96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not.

Fisher, Ronald (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd. p. 46. ISBN 978-0-05-002170-5.

Ronald Fisher

- It is the probability of the null hypothesis being false
- The smaller the p-value the more unlikely the H_o is true and the more likely the measured event will occur (Fisher set $p<0.05$)
- If p-value is large: then there is not enough evidence to reject H_o so result is statistically not significant
- Probably the most mis-used statistic ever

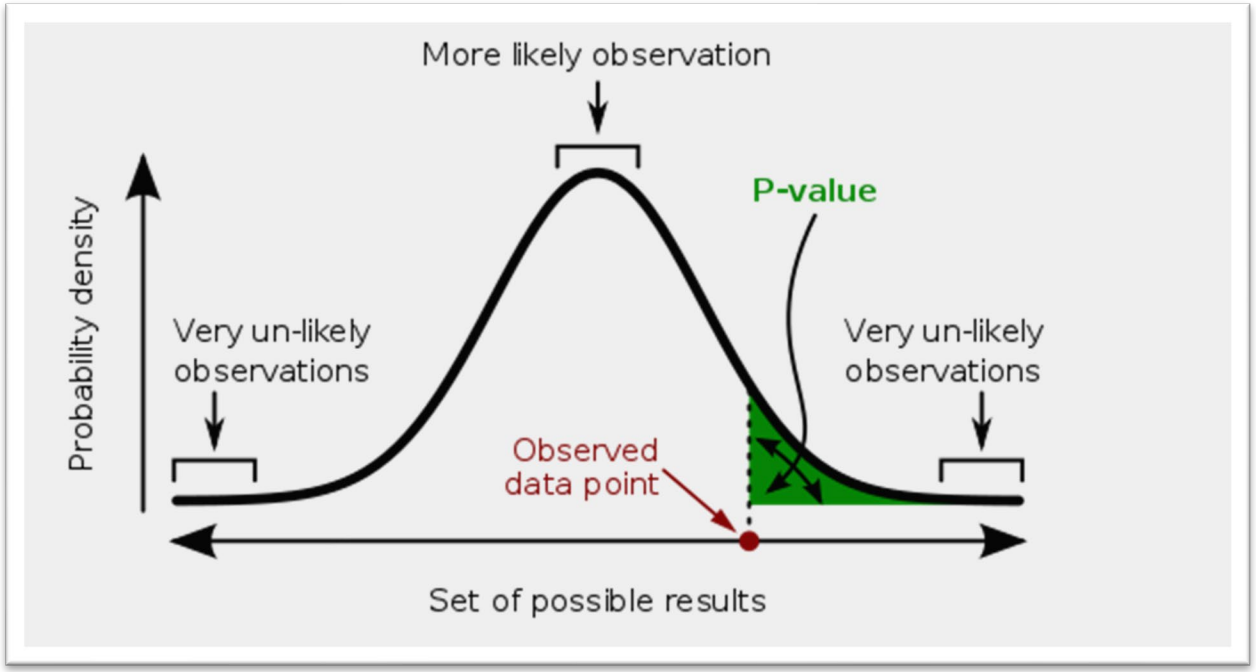
ASA Statement on p-Values

"the widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process"

Wasserstein, Ronald L.; Lazar, Nicole A. (2016-04-02). "The ASA's Statement on p-Values: Context, Process, and Purpose". *The American Statistician*. **70** (2): 129–133.

Statistical Significance

- $P<\alpha$
- α is usually set at 0.05
- 100% Arbitrary (Fisher)
- It is for normal distribution with 2 tails



User:Repapetilto @ Wikipedia & User:Chen-Pan Liao @ Wikipedia

Type I and II Errors

	H_0 is true Truly no difference	H_0 is false Truly a difference
Accept null hypothesis	Right decision	Wrong decision Type II Error (β)
Reject null hypothesis	Wrong decision Type I Error (α)	Right decision

Type I or alpha error is a false positive result

Type II or beta error is a false negative result

Avoid by having the right study group size – power analysis

Power

- Probability of demonstrating a statistically significant difference given there is a difference between the experimental and control groups
- Equal to 1-type II error
- Based on the magnitude of observed effect and the sample size of the study
  - The larger the effect, the smaller the sample size
  - The smaller the effect, the larger the sample size
- Must be done a priori (before collecting the data)
- Calculating the minimum sample size required to detect an effect
- Avoid a false negative result

Power Calculation – Sample size

A. Study Group Design

2 independent study groups or 1 study group vs population data

B. Primary end points

Dichotomous (yes or no) or Continuous (means)

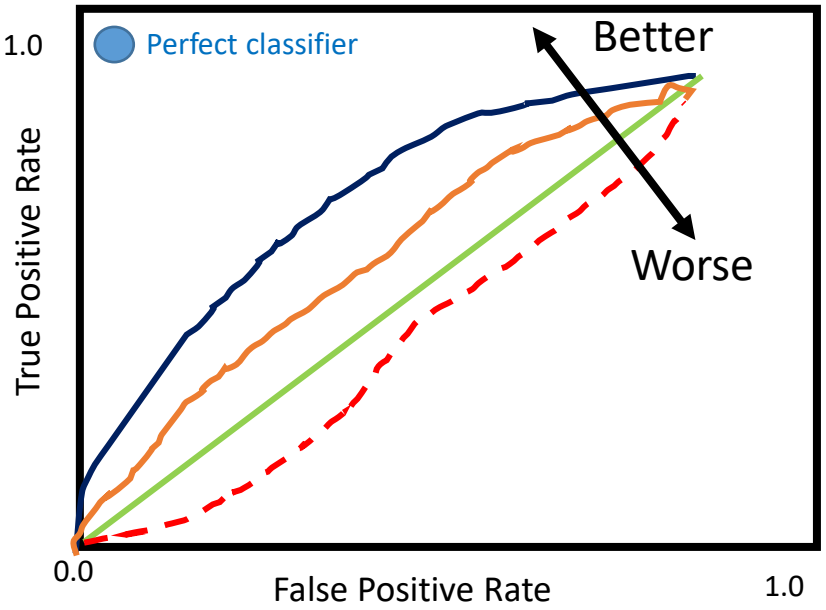
C. Statistical Parameters

Anticipated Outcome		Anticipated Error Rate	
Known Population/Group1	%	Type 1 alpha error rate	0.05%
Group 2	%	Power	80%

https://clincalc.com/stats/samplesize.aspx

Receiver Operator Curve (ROC)

A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

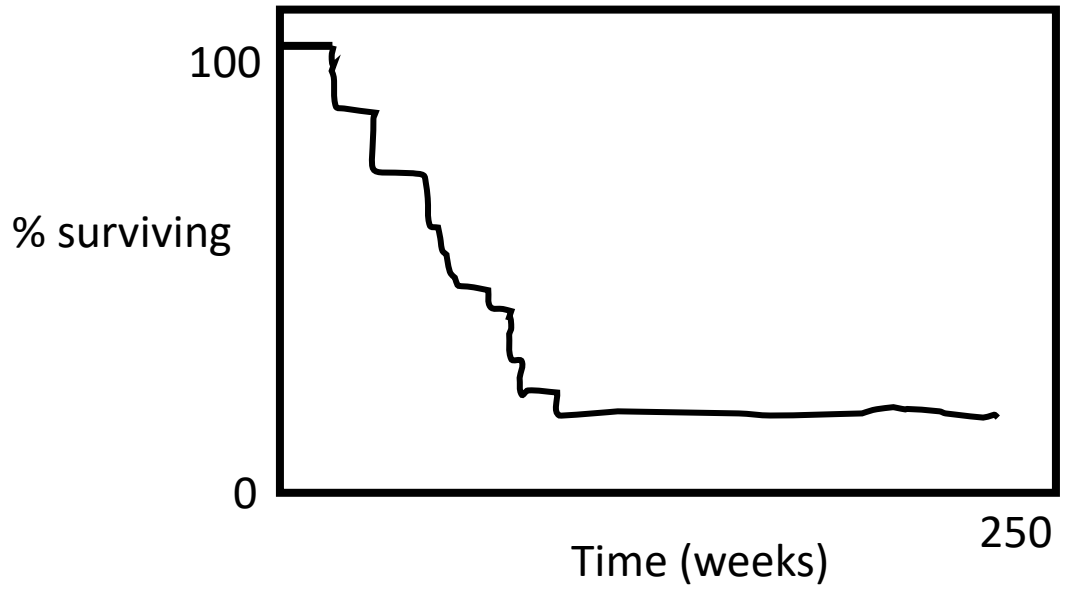


Area under the Curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

Survival Curve / Kaplan Meier

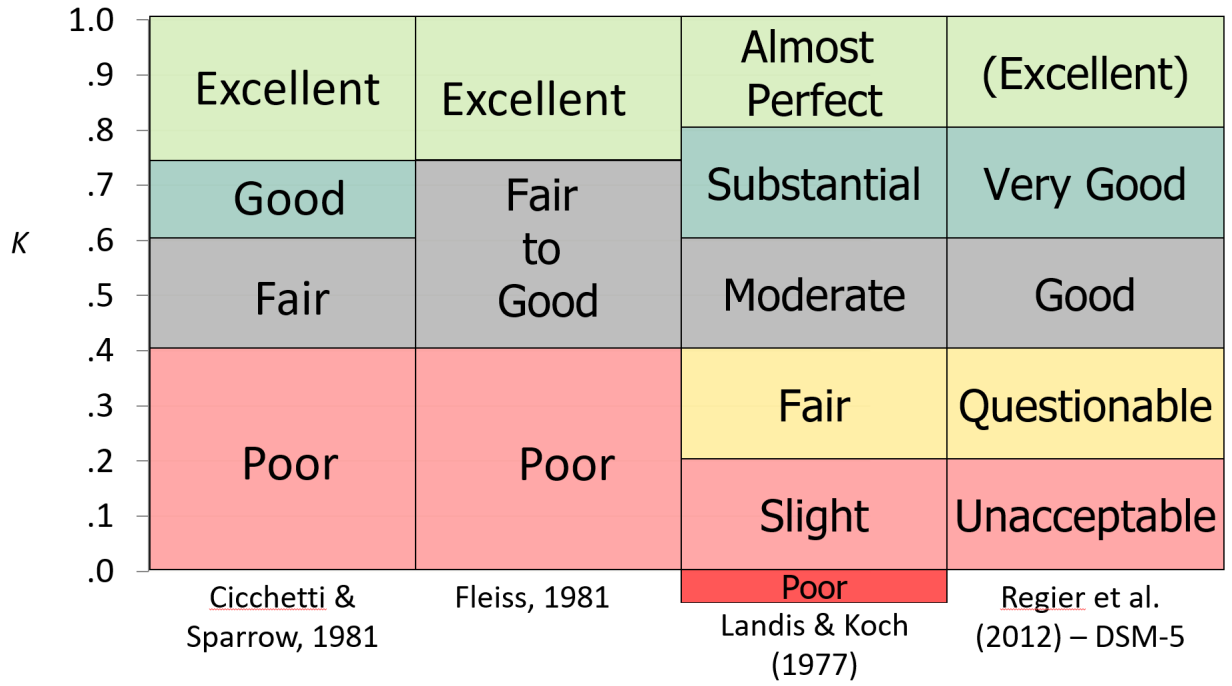
The visual representation of this function that shows the probability of an event at a respective time interval.

Survival probability is calculated as the number of subjects surviving divided by the number of patients at risk.



Cohen's Kappa Coefficient - Inter-Observer Reliability

- Kappa (κ)
- 0 = very poor agreement
- 0.6 considered reasonable
- A relative measurement



Sensitivity and Specificity – Predictive Values

- Proportions of positive and negative results in statistics and diagnostic tests that are true positive and true negative results,

- Positive Predictive Value

$$\frac{\text{Sensitivity X Prevalence}}{\text{sensitivity X prevalence + (1 – specificity) X (1- prevalence)}}$$

- Negative Predictive Value

$$\frac{\text{Number of true negative}}{\text{Number of true negatives + Number of false negatives}}$$

Study Design

Randomized Controlled Trial

- Gold Standard
- Must have a therapeutic uncertainty or equipoise
- Many confounders in surgery
- Many obstacles in surgery
- Narrow definition of participants may make the results not clinical useful to wide range of patients

Analytical Observational Studies

- Cohort: all patients with a specific intervention are studied
  - Prospective: define study parameters and enter patients as that intervention is applied and then followed for defined time period (looking forward)
    - More reliable, no data loss
  - Retrospective: define intervention that you wish to study as well as parameter and then review patients who have the intervention (looking back)
    - Less reliable, loss of data
- Case-Control: matched patients
  - Retrospective
  - Good for rare outcomes when large patient population is not available
  - Collect all the patients with a problem whether they have had the intervention or not and then match the non interventional with interventional patients

Descriptive Observational Studies

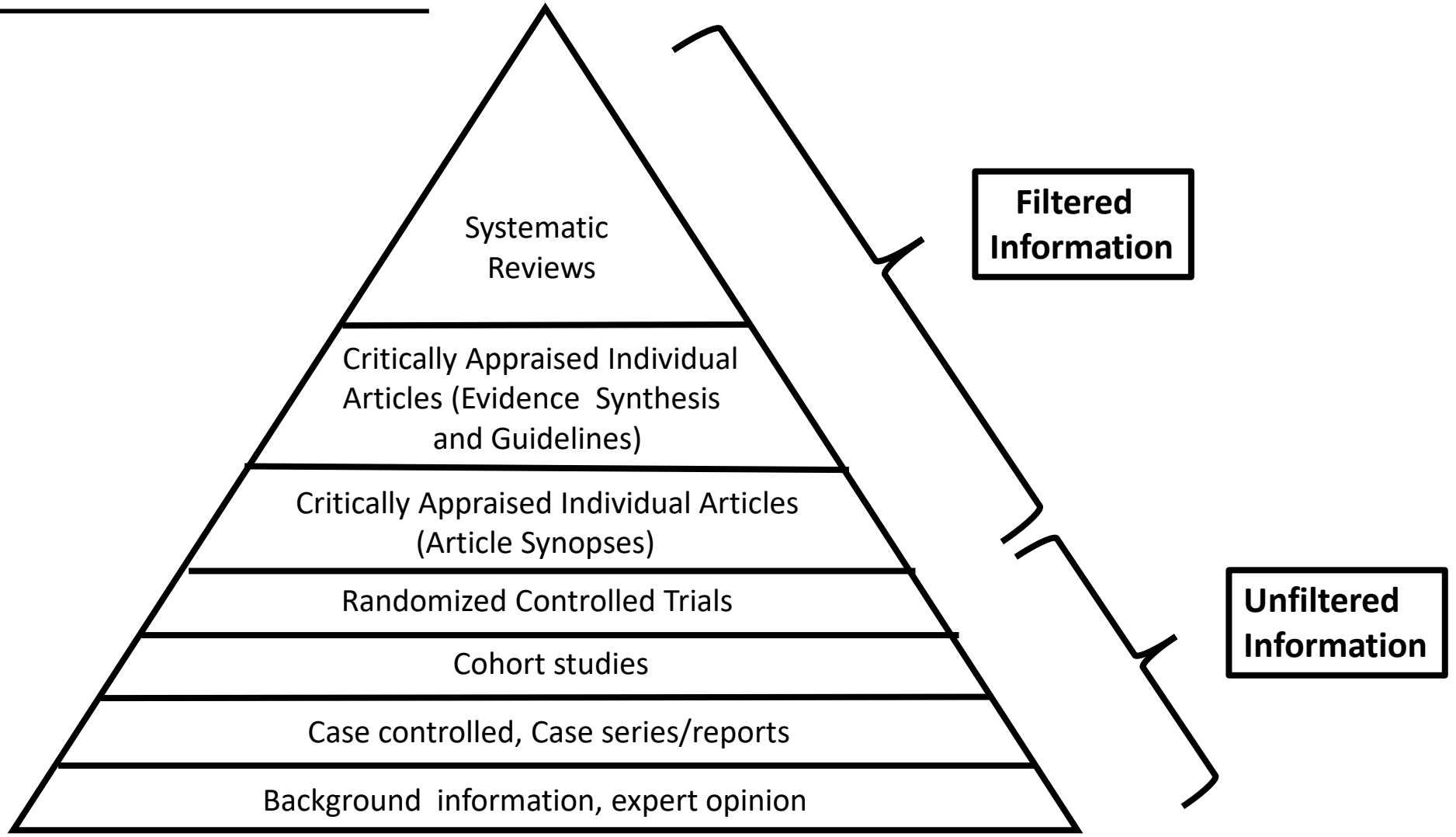
- Case Report
- Case Series

Levels of Evidence

- There are a number of systems
- Tool for assessment of results
  - Needs context
- JBJS levels of evidence
  - Level 1 – randomized controlled trial
  - Level 2 - prospective cohort or observational study with dramatic effect
  - Level 3 – retrospective cohort or case-controlled study
  - Level 4 – case series, historical controlled study

https://journals.lww.com/jbjsjournal/Pages/Journals-Level-of-Evidence.aspx

Pyramid of Evidence



https://journals.lww.com/jbjsjournal/Pages/Journals-Level-of-Evidence.aspx

Statistics gone wrong

- Choice of test – not always the *t*-test, must know the data type, your data distribution etc.
- Sources of bias and variation: channeling effect, publication bias, recall bias, data completeness bias
- Multiple Comparisons: need to understand how to calculate P-value
- Oversimplification of analysis: 69% of studies basic parametric tests
- Inadequate reporting:37% inadequate reporting of numbers,20% introduced new statistics in results section, 23% no measurement of error for main outcome, 16% reported different numbers in Methods and Results
- 51.5% of studies misused statistics and in a BMJ Study – 89% unacceptable at submission, down to 16% at publication

Statistics misunderstood

Poor review at the journal level (1998)

- 52% - 27% chance of pre acceptance stats review
- 31-82% chance of stats consultant on staff
- 50% of time stats review led to an important change

Goodman et al. J Gen Int Med 1998

Poorly Reviewed (20 years later)

- 34% do not use stats review
- 23% for all articles
- 24% in between
- >50% of reviews led to important changes

Hardwicke, Goodman, PLOS ONE Oct 1 2020

Summary

- Why statistics are important
- Basic statistical terms and test
- Common statistical missteps

References

Biostatistics: The Bare Essentials, Norman, Geoffrey R.; Streiner, David L.

Statistics and Data Management, A Practical Guide for Surgeons, D Stengel, M Bhandari, B Hanson. Thieme 2009

Clinical Epidemiology and Biostatistics: A Primer for Orthopedic Surgeons, M Kocher and D Zurakowski J Bone Joint Surg. Am. 2004; 86: 607-620

Thank You