Introduction to Statistics

Roger J. Lewis, MD, PhD
Department of Emergency Medicine
Harbor-UCLA Medical Center
Torrance, California

– SOMETIMES I HAVE TO GO THROUGH MANY DIFFERENT STATISTICIANS TO GET THE RIGHT RESULTS.

Classical Hypothesis Testing: Introduction

- A statistical plan or method for deciding which of two hypotheses is best supported by the data
- Uses a *p* value as the measure of the strength of evidence against one of the hypotheses

Classical Hypothesis Testing: The Null Hypotheses

- The hypothesis that there is no difference between the two groups to be compared, with respect to the measured variable
- Must be defined prior to data collection
- Must pass the "so what" test

Classical Hypothesis Testing: The Alternative Hypothesis

- The hypothesis that there is a difference between the two groups to be compared, with respect to the measured variable
- The size of the difference should be defined prior to data collection

Classical Hypothesis Testing: The Alternative Hypothesis

- The difference defined by the alternative hypothesis is usually the minimum clinically significant difference
- A larger difference is sometimes sought, if detecting the minimum clinically significant difference would require too large a study

Classical Hypothesis Testing: The *p* Value

- The null hypothesis is "tested" to determine which hypothesis (null or alternative) will be accepted as true
- Calculate the probability of obtaining the results observed, or results more inconsistent with the null hypothesis, if the null hypothesis were true
- This probability is the *p* value

Classical Hypothesis Testing: Rejecting the Null Hypothesis

- If the *p* value is less than some predetermined value, α, then the null hypothesis is rejected
- If the null hypothesis is rejected, then the alternative hypothesis is accepted as true
- Note that the alternative hypothesis is not directly tested

Classical Hypothesis Testing: Steps

1. Define the null hypothesis
2. Define the alternative hypothesis
3. Calculate a *p* value
4. Accept or reject the null hypothesis
5. Accept the alternative hypothesis if the null hypothesis is rejected

Classical Hypothesis Testing: Type I Error

- Concluding that a difference exists when it does not
- A false positive
- Occurs when a statistically significant *p* value ($p < \alpha$) is obtained when the two groups are not different
- The risk of a type I error, assuming there is no underlying difference, is α

Classical Hypothesis Testing: Type II Error

- Concluding that a difference does not exist, when a difference equal to the alternative hypothesis does exist
- A false negative
- Occurs when a *p* value > α is obtained, yet the two groups are different
- The risk of a type II error, assuming there is a difference, is β

Classical Hypothesis Testing: Power

- The chance of obtaining a statistically significant *p* value, if a true difference exists that is equal to that defined by the alternative hypothesis
- Power = 1 - β
- Power is determined by sample size, the magnitude of the difference sought, and by α

Steps in Sample Size Determination

1. Define the type of data (continuous, ordinal, categorical, etc.)
2. Define the size of the difference sought
3. Define α, the maximum significant *p* value
4. Determine the power desired (usually 0.80 or 0.95)
5. Look up the sample size in tables, or use published formulas or software

Statistical Tests

Test	Comparison	Principal Assumptions
Student's t test	Means of two groups	Continuous variable, normally distributed, equal variance
Wilcoxon rank sum	Medians of two groups	Continuous variable
Chi-square	Proportions	Categorical variable, more than 5 patients in any particular "cell"
Fisher's exact	Proportions	Categorical variable

Statistical Tests (Continued)

Test	Comparison	Principal Assumptions
One-way ANOVA	Means of three or more groups	Continuous variable, normally distributed, equal variance in all groups
Kruskal-Wallis	Medians of three or more groups	Continuous variable

Parametric vs Non-Parametric Tests

Parametric Test		Non-Parametric Test
Student's t test	→	Wilcoxon rank sum
One-way ANOVA	→	Kruskal-Wallis
Pearson correlation	→	Spearman rank correlation

Confidence Intervals: Example

- Purpose: to compare the effects of vasopressor A (V_A) and vasopressor B (V_B) based on post-treatment SBP in hypotensive patients
- Endpoint: post-treatment SBP
- Null hypothesis: Mean SBP_A = Mean SBP_B
- Results: Mean SBP_A = 70 mm Hg (after V_A)
  Mean SBP_B = 95 mm Hg (after V_B)
  Observed difference = 25 mm Hg ($p < 0.05$)
  25 mm Hg is the "point estimate"

Limitations of the *p* Value

- $p < 0.05$ tells us that the observed treatment difference is "statistically significantly" different than zero
- $p < 0.05$ does not tell us:
  - The uncertainty in the size of the true treatment effect
  - The likelihood that the true treatment effect is clinically important

The Point Estimate and the CI

- When using CIs, we would report the point estimate and the limits of the CI surrounding the point estimate, for example: 25 mm Hg (95% CI 5 to 44 mm Hg)

The Point Estimate and the CI

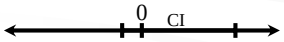
- Together, the point estimate and CI tell us:
  - The statistical significance of the difference (does the CI include zero?)
  - The size of the observed treatment effect
  - The uncertainty in the size of the true treatment effect
  - The likely clinical importance of the true treatment effect

Interpretation of the CI

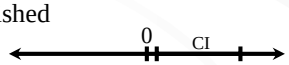
- Even if the data did not show a statistically significant difference, the CI can tell us if:
  - There probably really isn't a clinically-important difference between the treatments; or
  - There were not enough patients to reliably detect a clinically-important difference even if it really exists

Interpretation of the CI

- Even if the CI includes 0, if it also includes clinically important values, then potential benefit has not been ruled out



- Even if the CI does not include 0, if it includes clinically unimportant values then benefit has not been unequivocally established

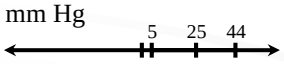


Interpretation of the CI

- Consider the comparison of vasopressor A and vasopressor B
- A difference of 0 is the null hypothesis
- Since the 95% CI, 5 to 44 mm Hg doesn't include 0, this is equivalent to $p < 0.05$
- Remember that for an odds ratio (OR) or a relative risk (RR) a value of 1 is equivalent to no difference

Interpretation of the CI

- Although the point estimate for the difference is 25 mm Hg, the results are consistent with the true difference being anywhere between 5 and 44 mm Hg



- Based on our own judgement of the minimum true difference that justifies a change in clinical practice, considering side effects, cost, etc., this may or may not justify a change in practice

Why a 95% CI?

- The selection of 95% CIs (as opposed to 99% CIs, for example) is arbitrary, like the selection of 0.05 as the cutoff for a statistically significant *p* value

Multiple Comparisons

- When two identical groups of patients are compared, there is a chance (α) that a statistically significant *p* value will be obtained (type I error)
- When multiple comparisons are performed, the risk of one or more false-positive *p* values is increased
- Multiple comparisons include:
  - Pair-wise comparisons of more than two groups
  - The comparison of multiple characteristics between two groups
  - The comparison of two groups at multiple time points

Multiple Comparisons: Risk of ≥ 1 False Positive

Number of Comparisons	Probability of at Least One Type I Error
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
10	0.40
20	0.64
30	0.79

Assumes α= 0.05, uncorrelated comparisons

Multiple Comparisons: Bonferroni Correction

- A method for reducing the overall risk of a type I error when making multiple comparisons
- The overall (study-wise) type I error risk desired (e.g., 0.05) is divided by the number of tests, and this new value is used as the α for each individual test
- Controls the type I error risk, but reduces the power (increased type II error risk)

Multiple Comparisons: Tests for Three or More Groups

- Analysis of Variance (ANOVA)
- Kruskal-Wallis test
- Chi-square test
- Fisher's exact test

$\Rightarrow$ These tests do not use the Bonferroni correction; they test the hypothesis that all groups are the same, and they preserve power

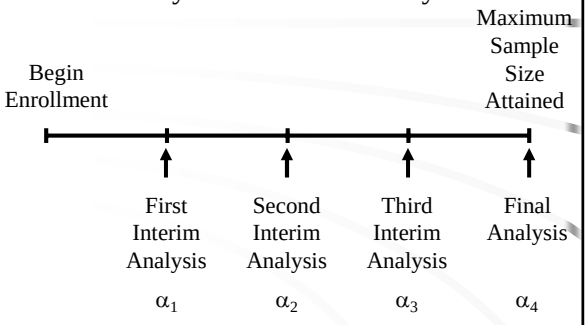
Interim Data Analyses: Ethical Motivation

- During a clinical trial, data accumulate sequentially
- If you were the last patient to be enrolled, wouldn't you want to know the treatment assignments and outcomes of the prior patients?
- Interim analyses are used to see if a difference clearly exists between the two groups, so the trial can be stopped early, and future patients can receive the best treatment
  - In other words, to stop the trial as soon as a reliable conclusion can be drawn from the available data

Interim Data Analyses: Statistical Considerations

- Interim data analyses are a type of multiple comparison
- Interim analyses must be planned in advance, including the amount of type I error risk to be taken at each analysis
- Large studies and studies of diseases with high morbidity and mortality should include planned interim analyses

Group Sequential Trial with Three Interim Analysis and a Final Analysis



Nominal α Levels

- α values (the maximum significant *p* value) for each interim analysis are adjusted downward, so that the true type I error rate for the entire study is 0.05
- Different patterns of nominal α values can be used:
  - Pocock design: constant α values
  - O'Brien-Fleming design: larger α values as trial progresses
    - Greater power for a given maximum N
    - More conservative at the beginning

Max No. Groups	Analysis	Pocock α_i	O'Brien-Fleming α_i	
2	Interim 1	.0294	.0052	
	Final	.0294	.0480	
3	Interim 1	.0221	.0005	} > 0.05
	Interim 2	.0221	.0141	
	Final	.0221	.0451	
4	Interim 1	.0182	5E-5	
	Interim 2	.0182	.0042	
	Interim 3	.0182	.0194	
	Final	.0182	.0430	

Subgroup Analysis: Motivation

- Patient populations are heterogeneous, composed of subgroups
- This is especially true for populations of emergency department patients
- A treatment effect detected in the entire population may or may not exist for a particular subgroup
- Data from subgroups are often clinically important and analyzed separately

Subgroup Analysis: Problems

- Analysis of multiple subgroups requires the use of multiple comparisons, increasing the overall risk of a type I error
- Since each subgroup is smaller than the whole study population, the power of subgroup comparisons is smaller, increasing the risk of type II error
- These problems occur even if the subgroups were defined prior to data collection

Subgroup Analysis: Problems

- Proper subgroup: Defined by characteristics available at enrollment, or which are not modified by the treatments being compared
- Improper subgroup: Defined by characteristics that can, in principle, be affected by study procedures or the treatments being compared
- Many retrospective studies include comparisons of improper subgroups (e.g., subgroup with "refractory shock")

James Stein Effect and Subgroups

- Even if the treatment works equally well in all subgroups, there will tend to be a "spread" in the apparent treatment effect when we analyze the data
- Similarly, the sizes of treatment effects are too "spread out" when we analyze the effect in each subgroup separately
- This is the James Stein effect

The James-Stein Estimator

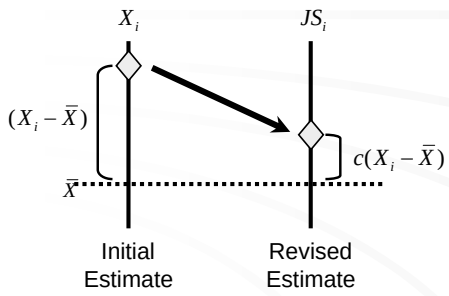
- Naïve approach: simply calculate the actual differences in outcomes, sound sophisticated by calling these the maximum likelihood estimates of the treatment effects, and use these values as the estimates
- James-Stein estimator:

$$JS_i = \bar{X} + c(X_i - \bar{X}), \text{ where } c < 1$$

- Best estimates are "shrunk" towards the group average

Efron B, Morris C. Stein's paradox in statistics. Scientific American 1977;236:119-127.

"Shrinkage"



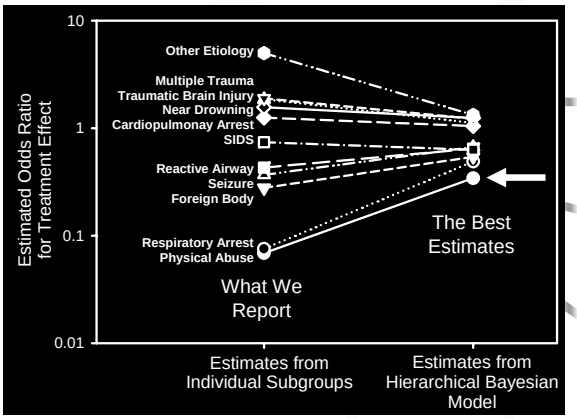
Treatment Estimates in Subgroups

- The best estimate of the true treatment effect in a subgroup is not the treatment effect observed in that subgroup, if there are 3 or more subgroups
- The James-Stein estimator was discovered 50 years ago, and yet we continue to report naïve estimates of treatment effects in subgroups

Example: Pediatric Airway Study

- Comparative trial of endotracheal intubation (ET) and bag-valve-mask ventillation (BVM) in the prehospital treatment of critically ill children
- Primary outcome: survival to hospital discharge
- Overall result: no improvement in survival
- Some evidence of harm and some evidence of benefit in clinically important subgroups, defined *a priori*

Gausche M, et al. Effect of out-of-hospital pediatric endotracheal intubation on survival and neurological outcome: a controlled clinical trial. JAMA 2000;283:783-790.



Determinants of Efficacy

- The effectiveness or efficacy of a therapy is determined by:
  - one's ability to administer the therapy to the patient or to get the patient to take the medication (i.e., "compliance")
  - inherent or "chemical efficacy"
  - other patient characteristics that you may not be able to anticipate, measure, or control

Compliance, Prognosis, and Bias

- Compliant and noncompliant patients often differ in many characteristics, including prognosis
- Even in a randomized, double-blind study compliance is rarely equal in the different treatment groups
- This can potentially introduce bias, in that the non-compliant, poor-prognosis (or good-prognosis) subgroup will tend to leave one treatment more than the other

Intention-to-Treat Analysis: Motivation

- To estimate the effectiveness of a treatment in clinical practice, one must properly allow for differences in compliance
- This is the purpose of the intention-to-treat principle

Intention-to-Treat Analysis: Definition

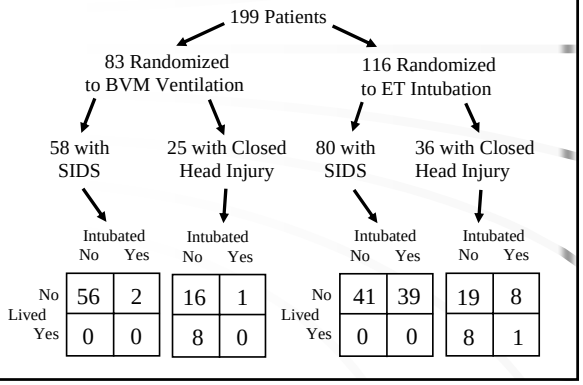
- Patients are considered to be members of the treatment group to which they are originally assigned, regardless of whether or not they receive that therapy
- In other words, patients are assigned to treatment groups according to the treatment they were intended to receive

Analysis by Treatment Received

- Patients are considered to be members of treatment groups based on what treatment they actually received
- Thus a patient assigned to an active drug treatment, who freely admits to ever taking any tablets, would be considered a member of the control group

Don't do this!

Example: Pediatric Airway Management



Intention-to-Treat Analysis: Example

- Intention-to-treat Analysis:
  - Survival in ET group: 7.8% (9/116)
  - Survival in BVM group: 9.6% (8/83)
- Analysis by treatment received:
  - Survival in ET group: 2.0% (1/51)
  - Survival in BVM group: 10.8% (16/148)
  - → Study would conclude that ET kills!
- Analysis by treatment received is misleading if there is a correlation between compliance and prognosis

Using Statistical Consultants: Guidelines (My Wish List)

- Define the most important question to be answered by the proposed study, in terms of measurable quantities
- For a comparative study: Define the size of the difference you wish to detect
- For an observational study: Define the precision with which you wish to measure the most important outcome

Using Statistical Consultants: Guidelines (Continued)

- Get as much information as possible about what you expect in the control group
- Define values for α and power, and the maximum sample size that is realistic
- Define clinically important subgroups of the population
- Determine whether there are multiple important comparisons

Using Statistical Consultants: Guidelines (Continued)

- Bring examples of published studies that illustrate the type of analysis you would like to perform at the end of the study
- Consider the feasibility of performing planned interim analyses of accumulating data