Common Statistical Issues in Biomedical Research

Howard Cabral, Ph.D., M.P.H.
Boston University CTSI
Boston University School of Public Health
Department of Biostatistics
May 15, 2013

- Overview of Basic Statistics
- Sample size and Power
- Types of Study Design

Overview of Basic Statistics

- Descriptive Statistics

Provide summaries of data and may be pictures, such as histograms, or plots, or numbers, such as counts, rates or means.

- Statistical Inference

Inferences about populations based on samples in which population **parameters** are estimated by sample **statistics.**

  - Estimation, e.g., Association
  - Hypothesis Testing

Describing distributions of variables

Measures of central tendency

- Mean
- Median
- Mode

Describing distributions of variables (cont.)

Measures of dispersion

- Standard deviation
- Variance
- Standard Error
- Range (maximum – minimum)

Describing distributions of variables (cont.)

Measures of rates

Proportion with event of interest

Statistical Inference

Estimation: Measures of Association

- For example, with "event/no event" data, let $\hat{p}_e$ be the estimated incidence rate in "exposed" group and $\hat{p}_{\bar{e}}$ be this rate in "unexposed" group

Statistical Inference (cont.)

- Relative risk (RR): $\frac{\hat{p}_e}{\hat{p}_{\bar{e}}}$
- Odds Ratio (OR): $\frac{\hat{p}_e/(1-\hat{p}_e)}{\hat{p}_{\bar{e}}/(1-\hat{p}_{\bar{e}})} = \frac{\hat{p}_e(1-\hat{p}_{\bar{e}})}{\hat{p}_{\bar{e}}(1-\hat{p}_e)}$
- Risk Difference (RD): $\hat{p}_e - \hat{p}_{\bar{e}}$

Statistical Inference (cont.)

95% confidence interval

- In large samples, a two-sided interval for a sample estimate, $(\hat{\theta})$, is:

$$(\hat{\theta}) \pm z_{\alpha/2} \times S.E.(\hat{\theta})$$

Hypothesis Testing

	H_0 false	H_0 true
H_0 rejected	correct	Type I error
H_0 not rejected	Type II error	correct

α = prob (Type I error)
β= Prob (Type II error)
power = 1 - β

Hypothesis Testing (cont.)

- Ideally, both α and β are small.
- Convention: choose α and use n large enough to keep β small.

Hypothesis Testing (cont.)

- Example: Two drugs are to be compared in a clinical trial for use in treatment of a disease. Drug A is cheaper than Drug B. Efficacy is measured using a continuous variable, y, and $H_0 : \mu_A = \mu_B$ versus $H_1 : \mu_A \neq \mu_B$.

- A Type I error occurs if they are truly equally effective but we conclude that, say, Drug B is better. The consequence is financial loss.

- A Type II error occurs if, say, Drug B is truly more effective but we fail to reject the null hypothesis, and conclude that they are equally effective. What is the consequence?

Hypothesis Testing (cont.)

- **"A p-value is not a 'level' "**

- **P-value**: probability that is **a function of sample data** and is defined as:

  Prob(test statistic as large as observed or larger | H_0 true)

Hypothesis Testing (cont.)

- Some prefer confidence intervals to P-values because they at least give a sense of the sample-based variability around the estimated effect in the units measured in that effect.
- A P-value is a unitless measure and is difficult to compare from analysis to analysis. P-values and confidence intervals, however, provide different information and are descriptive statistics.

Hypothesis Testing (cont.)

- **Alpha-level: predetermined criterion for** rejection of null
- **Power**: Failure to reject does not equal "acceptance" of null (Does “not guilty” mean “innocent” in a court of law?).
- Beware of a power problem when a result is of an important magnitude but is statistically not “significant”.

Other common misunderstandings

- "significant": a relative statement. One should ask, "What is your α?"
- One-sided (directional) vs. Two-sided tests (non-directional)

Sample size and Power

From an *ethical* perspective:

Too *few* subjects:

- Cannot adequately address the study question. Time, discomfort and risk to study subjects have served no purpose. This is unethical.
- May conclude that there is no therapeutic advance that is truly beneficial. Current and future subjects may never benefit from therapy based on inconclusive study. This is also unethical.

Sample size and Power (cont.)

Too *many* subjects:

- Too many subjects unnecessarily exposed to risk. Should enroll only enough patients to answer the study question.

Sample size and Power (cont.)

Before we can determine sample size, we need to answer the following:

- What is the principal measure of patient outcome?
- How will the data be analyzed to detect a treatment difference?
- How small a difference is clinically important to detect?

Sample size and Power (cont.)

Example: *Does the ingestion of large doses of vitamin A in tablet form prevent breast cancer?*

- Suppose we know from Connecticut tumor-registry data that the incidence rate of breast cancer over a one year period for women aged 45 – 49 is 150 cases per 100,000

Sample size and Power (cont.)

Women are randomized to:

- *Group 1*: Control group given placebo pills by mail. This group is anticipated to have the same disease rate as the registry (**150 cases per 100,000**)

- *Group 2*: Intervention group given vitamin A supplements by mail. The investigator anticipates a 20% reduction in risk (**120 cases per 100,000**)

Sample size and Power (cont.)

Design issues affecting sample size calculations

- Outcome measures: e.g., continuous or dichotomous
- Alternative hypothesis: 1-sided or 2-sided
- Detectable difference or clinically important difference
- Patient variability
- Desired α and β
- Allocation ratio
- Drop out rate

Sample size and Power (cont.)

SAMPLE SIZES
FOR SELECTED VALUES OF P_1, P_2, α AND β

P_1	P_2	Δ (P_1-P_2)	α	β	n (in each group)
.40	.15	.25	0.05	0.2	49
.40	.20	.20	0.05	0.2	82
.25	.05	.20	0.05	0.2	59
.30	.15	.15	0.05	0.2	121
.25	.125	.125	0.05	0.2	152
.125	.25	-.125	0.05	0.2	152
.125	.25	-.125	0.10	0.2	120
.125	.25	-.125	0.05	0.1	203
.125	.25	-.125	0.05	0.25	135
.15	.25	-.10	0.05	0.2	250

(2 SIDED TEST, EQUAL ALLOCATION)

Sample size and Power (cont.)

- Sample size is very sensitive to values of Δ.
- Large numbers are required if we want high power to detect small differences.
- Consider
  - Current knowledge
  - Likely improvement
  - Feasibility – available accrual

Sample size and Power (cont.)

- Examine a range of values, i.e., for several Δ, power find the required sample size; for several n, Δ, find the power.

- Need to increase sample sizes to account subjects lost to follow up.

Types of Study Design

Retrospective vs. Prospective studies

Prospective cohort study (Follow-up study)

- Classify pregnant women as "exposed" to a drug or not and follow them over time to observe birth outcomes.
- Framingham Heart study

Types of Study Design (cont.)

Retrospective cohort study

- Determine the exposure histories of "at risk" workers at a shipyard over time and relate them to their disease histories.
- Cheaper than prospective cohort design

Types of Study Design (cont.)

Case control study

- Select a study population of children with anomalies and some without, then determine their exposure histories.
- Used for rare outcomes.

Types of Study Design (cont.)

Randomized Controlled Trials (RCTs)

- Randomly assign subjects to study groups, including a control group, and follow until outcome.
- Randomization with large enough sample size ensures that potential confounding variables are equally distributed across study groups. If so, bivariate analyses, e.g., t-tests, chi-square tests, are only needed to compare groups with validity.

Multiple hypothesis testing: To control or not to control?

- For a given study, we would like there to be one question (or at least a limited number) defined *a priori* for which an hypothesis is stated and an alpha level is set.

- We then collect the data and perform an analysis that provides evidence with respect to the hypothesis.

Multiple hypothesis testing: To control or not to control? (cont.)

- For example, in a clinical trial with three groups (A, B, and C), we might be interested in comparing mean systolic blood pressure across the groups.

- We can analyze these data via analysis-of-variance. If we find that the groups differ significantly on the whole (globally) at the specified alpha level, the expected subsequent question would be, "Which groups significantly differ?"

Multiple hypothesis testing: To control or not to control? (cont.)

- If these comparisons are limited in number and can be defined in advance, performing separate t-tests (for example A vs. B, B vs. C; A vs. C in a 3 group study) may be of interest and is acceptable.
- In other instances, one might specify comparisons after having looked at the data. This could result in a large number of comparisons.

Multiple hypothesis testing: To control or not to control? (cont.)

- For example, with 10 groups in a study, there are $\binom{10}{2}$ =45 possible pairs of comparisons. With an alpha level of 0.05, we would expect

  0.05×45, or 2 comparisons, to be significant by chance alone.

Multiple hypothesis testing: To control or not to control? (cont.)

- Our goal, then is to maintain the overall alpha level for the study, adjusting downward the alpha level for individual *post hoc* comparisons so that in sum the *experimentwise error rate* is maintained.

- The Bonferroni correction is an example of such an adjustment.

Multiple hypothesis testing: To control or not to control? (cont.)

Bonferroni correction

- Set a new alpha called, α^*, so that the overall alpha is maintained such that $\alpha^* = \alpha / \binom{k}{2}$ for *k* pairwise comparisons.

- Without such a correction, the probability of <u>incorrectly</u> finding at least one significant comparison markedly increases.

Multiple hypothesis testing: To control or not to control? (cont.)

- Other well-known experimentwise adjustment methods include Scheffe's test, Tukey's test, and Dunnett's test.
- There is some controversy in the epidemiologic literature surrounding the issue of adjusting for multiple hypothesis testing. Points of criticism are made around the hypotheses underlying P-values, as well as concerns of practicality in large studies where many questions can be addressed.

Summary

Statistical issues inherent in a given research question are often more complicated than you realize. Consult with a statistician before you collect your data.
He or she will address issues such as:

- Power/sample size?
- Does the study design best address research question?
- Which analysis is best suited for the design? Often more than one is acceptable.

Howard Cabral
Crosstown 3rd floor, Room 310
8-5024
hjcab@bu.edu