Statistical Thinking in Biomedical Research

Department of Biostatistics
School of Medicine
Vanderbilt University

biostat.mc.vanderbilt.edu/biostat/twiki/bin/view/Main/ClinStat

OCTOBER 2, 2004

Section 1: Introduction to Statistical Concepts — Frank Harrell

- What does biostatistics have to offer to biomedical research?
- Statistical inference
- Study design issues
- Descriptive statistics
- Measuring change
- Biostatistical resources at VU

What Does Biostatistics Offer?

- Help in developing concrete objectives and data acquisition methods that meet the objectives, concrete descriptions of primary and secondary endpoints
- Appropriate experimental and study design
  - Sources of bias
  - Measurement issues
  - Efficiency/power
  - Maximizing use of a given number of animals
  - Interpretability of findings
  - Reproducibility of analyses

  Choice of appropriate design depends crucially on the type of experiment/disease and treatment being studied.
- $\uparrow$ likelihood that sample will yield estimates of adequate precision to make experiments

conclusive/affect medical practice

- More efficient use of data
- Formulate analysis plans without making inappropriate assumptions
- Estimate sample size (if fixed)

Example

- Objective: Does an intervention (could be a new drug, patient education intervention, method of administration) improve a response?
- What is meant by “response” and how will it be measured?
- Is it based on symptoms, physiologic measurements, anatomical measurements, or a combination?
- Does the “response” variable truly measure the effect of treatment?
- When is the response measured — 30 days after enrollment; 30 days after discharge from the hospital, anytime within a 5 year follow–up period, during the procedure, immediately upon reperfusion after a coronary artery is unclamped, after steady state, . . . ?

- If a “time to” endpoint, how much time is given to enrolling patients and how much to follow–up?
- Where will patients come from and which group do they represent?
- How will the study results be used?

Statistical Inference — Examples

- "How did my 5 patients do after I put them on an ACE–inhibitor?": Describe results.
- "How do patients with condition x respond after being on an ACE–inhibitor for 6 months?": Infer $\longrightarrow$ Need to take a sample of patients of interest to approximate what would be observed had *all* such patients been treated that way.
- "What is the in–hospital mortality after open heart surgery at my hospital so far this year?": Describe; whole population captured.
- "What is the in–hospital mortality after open heart surgery likely to be this year, given results from last year?": Infer $\longrightarrow$ Estimate probability of death for patients like those seen in 2003.

- Inference = observations $\longrightarrow$ some general truth
- Answering research questions usually requires inferential reasoning because you want to make a statement in general, not just a statement about your specific study.
- Ability to do so depends on how observations collected as well as their number

Infinite Data Case

- Suppose that one had an infinitely large amount of data of the kind under consideration
- Inference not required
- Do need to determine if the infinite dataset would answer the question of interest (QOI)
  - Subjects relevant?
  - Measurements biased?
  - Measurements relevant (e.g. measure cholesterol reduction but not survival time)?
  - Data collection process adequate?
  - Patient–to–patient variability still too great for conclusions to be applied to individuals?

Finite Dataset

- Compute an estimate of something, e.g. expected reduction in blood pressure
- Approximates what would have been observed if had ∞ data
- Can estimate likely $\left|\text{error}\right|$ in this approximation
- Probabilistic thinking: likely absolute error is a function of:
  - Sample size
  - Subject to subject variability
  - Intra–subject variability if using multiple observations/subject
  - Systematic bias
  - Some subjects not getting desired experimental condition

Steps Involved in Statistical Inference

- Statistical inference based on the fact that when laws of probability govern data collection, can infer from sample to infinite data results
- First insure that an infinite dataset would answer the QOI
- Assess results using a sample
- Compute likely closeness with which sample results approximates infinite dataset results
- Internal validity (chance), external validity (generalizability)

Study Design Issues

- Concretely define study objectives
- Design study so that if had ∞ data would answer QOI
- Conduct experiment making efficient use of resources, minimize likely $\left|\text{error}\right|$
- Standardize measurement devices
- Quantify and minimize intra– and inter–observer variability of measurements
- Define terms: symptoms, signs, diagnoses, disease severity, risk factors, treatment or experimental conditions, control condition, events
- Use standardized assessment instruments when possible
- Definite animal/patient entry criteria
- Concomitant therapies/laboratory conditions

- Dosing of active control agents must be optimal
- Account for accommodation (tolerance) to drug effect
- In comparison studies, masked and random assignment of experimental conditions
- Masked assessment of specimens/subject responses
- Masked specification of analysis[6]
- Masked reporting: write manuscript before data analysis[7]

Response Variables

- Continuous measurements are best (e.g., mmHg, not “hypertension”)
- Time lapse after experimental condition/how often to measure
- May have to wait until after an acute but temporary derangement
- Time of assessment may need to correspond with phases of disease development/trajectory of disease severity as well as lifespan of the technology
- Binary response when time of event not important (e.g., procedural death) — still need to justify duration of observation
- Time to event: all subjects without event need to be followed a minimum duration to capture some of the clinically relevant period.

Sufficiently large subset of subjects should be followed until the end of the clinically relevant period.

- Ordinal responses can be useful and have good statistical power, e.g., no event within 30d, mild myocardial infarction, moderate MI, severe MI, death within 30d. For diagnostic studies may need to at least include a "gray zone".
- When have multiple responses they should be (1) prioritized, or (2) combined into a summary scale. Need to "go out on a limb" and pre–specify which results will be emphasized when study results are publicized.
- Example: may combine systolic and diastolic b.p. into mean arterial b.p.
- Otherwise, have multiple comparison problems. To preserve overall type I error (false positive rate), would need to be more conservative $\longrightarrow \downarrow$ power.

Types of Studies/Believability of Results

- Single–arm (pilot, Phase I–II)
  - Comparisons with a reference standard
  - Toxicity
  - Pharmacokinetic
  - Correlating two responses
  - Dose–finding/dose titration
  - Estimation of dose– or time–response curves within subjects
- Beware of problems with noncomparative studies of therapies:
  Treatment response = natural history + Hawthorne effect + placebo effect + bias caused by investigator enthusiasm + real treatment effect
- Comparative: ≥ 2 arms
- Unacceptable Studies:
  - Observational studies where subjects were

selected on the basis of their outcomes (e.g., consecutive series of 100 open heart patients who lived)

- Comparison with historical controls unless time–trends fully understood and excellent subject baseline descriptors in both studies
- Randomized controlled trial (RCT) where physicians only allowed patients to be randomized who were invincible
- RCT where entry criteria otherwise do not reflect patients seen in practice
- RCT of a procedure or therapy that is obsolete by the time the results are disseminated (or mode of use is obsolete)
- Any study where positive results were derived only after torturing the data (multiple subgroups or response variables examined)

- Experiment in which measurements have extremely large variability across replications within the same animal, or ones in which measurements were “optimized” by non–replicable “tweaking”

- An average ranking of quality for comparative studies[a]:
  1. double masked RCT with masked analysis & manuscript writing
  2. double masked RCT
  3. single masked RCT
  4. unmasked RCT
  5. prospectively designed and conducted cohort study
  6. prospective case–control study
  7. retrospective cohort study
  8. retrospective case–control study
- See Chalmers et al.[4] for a rating scale for study quality. Also see[3].

[a] RCTs include crossover studies, which can be of excellent quality when there are no carryover effects or when carryover effects are understood well enough to be "subtracted out".

Randomized Experiments

- Randomly allocate patients to treatments while masked
- If sample size is at all reasonable, should balance all known and unknown risk factors
- Even if there is an apparent imbalance in one factor, you'll see imbalances in the other direction if you look at enough other factors
- Best not to look at patient characteristics stratified by treatment; report statistics for overall sample
- Randomization is best done using a computer program, with treatment assignments revealed at the last moment

External Validity of Study Findings

- Knowledge of pathophysiology can allow extrapolation of results to a group of subjects not represented in study
- Example: Reduction of probability of myocardial infarction by aspirin in men $\longrightarrow$ reduction in women
- **But** what if aspirin $\uparrow$ GI bleeding in women more than in men?
- Differences in dosing, side effects, compliance can cause different results in another population
- Relative effects of treatments frequently carry over to other types of subjects even though absolute effects do not[a]

[a] Because absolute risks of events vary with disease severity, dictating that risk differences must vary.

Pitfalls in Analysis & Interpretation

- Highlighting results found by data dredging; need to at least document the context
- Deleting "outliers" based on observed response values
  - Unscientific, results non–replicable
  - Instead use robust statistical methods
- Irreproducible analyses based on point–and–click software without audit trail
- Concentrating exclusively on hypothesis testing. Null hypotheses are generally boring and do not answer questions about clinical significance. It's better to think in terms of being able to estimate, with sufficient precision, the effects of interest.
- Using P–values to provide evidence supporting a hypothesis; they can only be used to quantify evidence *against* a hypothesis.

"Absence of evidence is not evidence of absence" [1]

- $P = 0.4 \rightarrow$ insufficient sample size or no effect; don't know which
- Relying too much on standard deviations as descriptive statistics. Standard deviations are not very meaningful if the distribution of the data is non–Gaussian and especially if asymmetric.
- Using standard errors to describe anything other than the precision of a summary estimate. Standard errors do not describe variability across subjects. To describe precision, it's better to use confidence limits on summary statistics.

Descriptive Statistics

- Number of non–missing measurements, central tendency, perhaps inter–subject variability
- Mean and especially standard deviation may not be meaningful unless data normally distributed
- Don't expect normality for biological variables
- Deciding on statistics to use on basis of test of normality assumes such tests have power near 1.0
- For continuous variables, a good summary is obtained from the 3 quartiles ($25^{th}, 50^{th}, 75^{th}$ percentiles, 50^{th} = median)
- Describes central tendency, spread, symmetry
- For continuous variables for which totals may be relevant (e.g., costs), supplement this with the sample mean
- Computing means on transformations (e.g.,

geometric mean) and then back–transforming is problematic

- Standard errors are not descriptive statistics
- For discrete numeric variables representing counts or interval scale values, where the number of possible categories < 10, use the mean and outer quartiles or mean and selected proportions. Median will not be sensitive and is erratic because of heavy ties in data.
- Nominal (polytomous) variables $\longrightarrow$ proportions in $k - 1$ of the k categories
- Binary variables $\longrightarrow$ mean (proportion of "positives")

Analysis of Paired Observations

- Frequently one makes multiple observations on same experimental unit
- Can't analyze as if independent
- When two observations made on each unit (e.g., pre–post), it is common to summarize each pair using a measure of effect $\longrightarrow$ analyze effects as if (unpaired) raw data
- Most common: simple difference, ratio, percent change
- Can't take effect measure for granted
- Subjects having large initial values may have largest differences
- Subjects having very small initial values may have largest post/pre ratios

What's Wrong with Percent Change?

- Depends on point of reference — which term is used in the denominator?
- Example:
  Treatment A: 0.05 proportion having stroke
  Treatment B: 0.09 proportion having stroke
  Treatment A reduced proportion of stroke by 44%
  Treatment B increased proportion by 80%
- Two increases of 50% result in a total increase of 125%, not 100%
- Percent change (or ratio) not a symmetric measure
- Simple difference or log ratio are symmetric

Objective Method for Choosing Effect Measure

- Goal: Measure of effect should be as independent of baseline value as possible[a]
- Plot difference in pre and post values vs. the average of the pre and post values. If this shows no trend, the simple differences are adequate summaries of the effects, i.e., they are independent of initial measurements.

[a] Because of regression to the mean, it may be impossible to make the measure of change truly independent of the initial value. A high initial value may be that way because of measurement error. The high value will cause the change to be less than it would have been had the initial value been measured without error. Plotting differences against averages rather than against initial values will help reduce the effect of regression to the mean.

- If a systematic pattern is observed, consider repeating the previous step after taking logs of both the pre and post values. If this removes any systematic relationship between the average and the difference in logs, summarize the data using logs, i.e., take the effect measure as the log ratio.
- Other transformations may also need to be examined

Biostat Resources at VU

- School of Medicine has decided that a strong biostatistics program and infrastructure is a priority
- Principal method for providing support to other divisions and departments: collaboration cost-sharing plan
  See `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/CollaborationProcedures`
  - Difficult for us to maintain unfunded percent efforts and availability of personnel for short-term consulting needs
  - We emphasize *collaboration* instead of emergency *consultation*
  - Collaboration cost-sharing plan: method for long-term integration of biostatisticians into your research program and to obtain high-priority assistance for grant proposals

  - Dept. of Biostatistics pays for 1/3 of PhD and MS biostatistician percent efforts devoted to non-grant-funded biomedical research development for **your** division in the School of Medicine
  - Describe your long-term needs at `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/OthDeptNeeds`

- Groups for whom we currently provide general non-grant-funded support (see `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/CollaborationAssignments`):
  - Department of Obstetrics & Gynecology
  - Department of Ophthalmology
  - Division of Orthopedic Trauma
  - Sports Medicine
  - Kennedy Center

- For Divisions not providing general ongoing support of a portion of an M.S. biostatistician and a faculty biostatistician, hourly charged assistance is on a first-come first-serve basis with a minimum of **60 days** advance notice before a grant application deadline. Availability of consultants is not guaranteed, due to demand and to prioritization of work for divisions and departments with whom we have long-term relationships. Also, consultants may not be biostatisticians with previous experience in your subject matter area.

- Who to contact:
  - General E–mail address: `biostat@vanderbilt.edu`
  - Claudia Calderon MBA
    Assistant to the Chair
    Department of Biostatistics
    claudia.calderon@vanderbilt.edu
    322-2001

How to Collaborate with Statisticians

- Willingness to explain the details of the study. An appropriate choice of the outcome, design, sample size, data collection, etc. requires some knowledge of the area being studied. Explaining these things can not only provide a more efficient way of doing the study, but can sometimes help to clarify issues that may have been taken for granted.

Collaboration Issues

- Collaboration should begin early
- Too often stat. will uncover a fatal flaw in data collection too late, e.g., recording measurements as ranges rather than raw data
- Early understanding on authorship; depends on whether e.g. statistician serves as a “number cruncher” vs. as part of the investigation or manuscript writing or she develops/assimilates new methods for the purpose of the project
- Best ways to fund biostat involvement are through the collaboration cost–sharing plan (at the divisional level) or through %FTE of grant support
- Long–term goal of Department of Biostatistics is to have stat. on staff who have long–term collegial relationships with biomedical researchers and with a good understanding of specific subject matter areas

Education Opportunities at VU

- MSCI program
- MPH program
- GCRC workshops
- Short courses (*Statistical Thinking* will be offered at least twice/year)
- Biostatistics seminars and workshops (see `biostat.mc.vanderbilt.edu`)

References

[1] D. G. Altman and J. M. Bland. Absence of evidence is not evidence of absence. *British Medical Journal*, 311:485, 1995.

[2] J. C. Bailar III and F. Mosteller. *Medical Uses of Statistics*. NEJM Books, Boston, second edition, 1995.

[3] C. Begg, M. Cho, S. Eastwook, R. Horton, D. Moher, I. Olkin, and *et al.* Improving the quality of reporting of randomized controlled trials. The Consort statement. *Journal of the American Medical Association*, 276:637–639, 1996.

[4] T. C. Chalmers, H. Smith, B. Blackburn, B. Silverman, B. Schroeder, D. Reitman, and A. Ambroz. A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, 2:31–49, 1981.

[5] T. J. Cole. Sympercents: symmetric percentage differences on the 100 $\log_e$ scale simplify the presentation of log transformed data. *Statistics in Medicine*, 19:3109–3125, 2000.

[6] CPMP Working Party. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. *Statistics in Medicine*, 14:1659–1682, 1995.

[7] P. C. Gøtzsche. Blinding during data analysis and writing of manuscripts. *Controlled Clinical Trials*, 17:285–293, 1996.

[8] L. Kaiser. Adjusting for baseline: Change or percentage change? *Statistics in Medicine*, 8:1183–1190, 1989.

[9] R. A. Kronmal. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society A*, 156:379–392, 1993.

[10] T. A. Lang and M. Secic. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. American College of Physicians, Philadelphia, 1997.

[11] J. S. Maritz. Models and the use of signed rank tests. *Statistics in Medicine*, 4:145–153, 1985.

[12] L. Törnqvist, P. Vartia, and Y. O. Vartia. How should relative changes be measured? *American Statistician*, 39:43–46, 1985.