Statistical Principles to Live By

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
f.harrell@vanderbilt.edu
biostat.mc.vanderbilt.edu

APPLIED STATISTICS 2004 LJUBLJANA, SLOVENIA
21 September 2004

Outline

- Nondescriptive Statistics
- Statistical Tests
- Respecting Continuous Variables and Avoiding Classification and Change Scores
- Choice of effect index
- Filtering Results

Outline, *continued*

- Underfitting and Overfitting
- Problems with Multi-stage Procedures
- Model-Building Strategies
- Do Simple Things Well
- Statistical Computing and Graphics

Nondescriptive Statistics

- $\bar{X}$ and SD virtually assume symmetry
  - $\bar{X}$ not representative of "typical" subject
  - SD difficult to interpret
- Let the data speak for themselves
- Three-number summary: 25th, 50th (median), 75th percentiles
- Describes central tendency, spread, symmetry

Nondescriptive Statistics, *continued*

- Don't say "mean cost was \$10,000 $\pm$ \$15,000"
- Routinely use the bootstrap for asymmetric confidence limits for μ

Nonparametric Tests

- Preferred if only want a P-value and situation is simple
- Power generally exceeds that of parametric tests
- Robust, transformation invariant
- Pre-testing for normality then choosing a test is a bad idea

Hypothesis Testing is Overused

- Often not interested in whether an effect is nonzero
- Usually interested in estimating magnitude of an unknown effect
- Confidence intervals or Bayesian posterior intervals preferred to P-values
- Avoid "exact" tests
  - Agresti: "The price of exactness is conservatism"

Adjustment for Multiple Comparisons

- Statisticians are good about multiplicity adjustments of P-values
- Not necessarily good about point estimates
  - E.g. huge bias in est. treatment effect if choose subgroup with smallest P-value
  - Gene microarray findings (gene expression ratios) overstated

Respecting Continuous Variables

- Keep all continuous variables continuous
- Huge loss of power and precision if dichotomize a continuous predictor or response variable
- Categorization assumes a discontinuous relationship
- Results in estimates applicable only to groups, not individuals
  - Risk of stroke for low vs. high blood pressure

Respecting Continuous Variables, *continued*

- Assuming linearity is better than assuming a piecewise flat relationship
- Better: nonparametric regression or parametric regression splines
- Recursive partitioning (& CART) make poor use of continuous predictors

Avoid Change Scores

- Change measure seldom checked for adequacy (data properly normalized)
- Better: analysis of covariance, adjusting for baseline value
- Predict final value, estimate changes later
- If change score used, baseline value must appear on both sides of model equation

Classification vs. Prediction

- Statisticians should provide predictions, not classifications
  - Probability of disease
  - Probability of survival past t
  - Life expectency
  - Predicted blood pressure at 2 months
- Leave classification up to the possessor of the utility function (usually not the analyst)

Choice of Effect Index

- Index should be symmetric (log ratio or ratio, not % change)
- Should be context–free
- Risk difference may be good for communicating to a patient but is not sufficient for communicating the results of an analysis
- Risk difference and ratio are not capable of being constant

- Odds and hazard ratios are

Avoid Filtering of Results

- Reporting only the one of many endpoints that was "significant"
- Subsetting data to find an effect
- Removing ineffective treatments from consideration
- Truncating follow-up time when late results make a treatment look bad
- Removing insignificant predictors from the model

Avoid Underfitting

- Unwarranted linearity assumptions
- Using t-test or ANOVA instead of ANOCOVA
  - In perfectly balanced randomized experiment with binary or time to event endpoint, failure to adjust for subject heterogeneity biases treatment effects towards null

Avoid Overfitting but Not Shrinkage

- Fitting model more complex than information content supports
- **Many** published models are overfitted; be skeptical
- Other authors attempting to validate a published model falsely assume non-transportability
- An unbiased validation would have revealed poor fit in the original analysis
- Use shrinkage (discounting, penalization)

Problems with Multi-Stage Procedures

- Few practicing statisticians know how to simulate to find true operating characteristics of such procedures

  Brownstone [1]: "theoretical statisticians have been unable to analyze the sampling properties of [usual multi-step modeling strategies] under realistic conditions." He concludes that the modeling strategy must be completely specified and then bootstrapped to get consistent estimates of variances and other sampling properties. See also [2, 3].

- Pre-testing for normality
- Pre-testing carryover effect in crossover studies (Senn)

Multi-Stage Procedures, *continued*

- Trying > 2 parametric distributions that best fit ECDF or Kaplan-Meier esimates
  - Parametric cumulative probability or quantile estimates inherit imprecision of nonparametric distribution estimates when properly compute variance of estimates
    var($\hat{\theta}|$model correct) low
    var($\hat{\theta}$) higher

Multi-Stage Procedures, *continued*

- Seeking "optimum" cutpoints for testing association, treating cutpoints are if pre-specified
- Univariable screening
- Stepwise variable selection

Problems with Extreme Flexibility

- Automatic interaction detection
- Recursive partitioning (& CART)
- Price of strictly empirical procedure not driven by science can be conservatism
  - Often must prune trees back until R^2 is low
- If additivity assumption is 0.6 correct, a flexible additive model can outperform recursive partitioning for commonly used N

Use Better Modeling Strategies

- Be unafraid of complex models; graph for non-statisticians
- Use a strategy you can program; can study properties by simulation
- Use subject matter knowledge more than P-values to guide model selection
- In ordinary situations, fitting full pre-specified model performs better than statistical variable selection

Model Strategies, *continued*

- Otherwise if there is model uncertainty it is better to average models than to select a single "winner"
- Data reduction and shrinkage have advantages
- Make transformation estimation a part of model fitting
  - Will get correct d.f., α, confidence intervals
- Avoid casewise deletion of missing data

Modeling Strategies, *continued*

- Validate model performance unbiasedly
  - Holding back test data from model development/fitting is inefficient
  - Mean squared error of accuracy estimate is high
  - Resampling techniques preferred
  - Must consider **all** aspects of model uncertainty
- When data mining, the weapon of mass destruction you find may not be the one you seek

Weapon of Mass Destruction

Do Simple Things Well

- Ordinary multiple regression is not well done by many statisticians
- Addressing nonlinearity is very important, and can be done simply
- Understand absolute effects
  - Dominated by background or control group risks
  - Severity of disease is very important
- Pharmacogenomics is unlikely to provide treatment

selection rules as effective as simple rules based on background risk that are currently available

- Individualized medicine should be largely based on simple ideas

Use Better Graphics

- Only pie charts are worse than bar charts (especially vertical ones)
- Dot charts have many advantages
- Box plots, extended box plots, ECDFs, rug plots, scatterplots, confidence bands, quantile bands are some of the many effective graphical devices
- Replace at least $\frac{1}{2}$ of tables with graphics

Use Modern Statistical Computing Methods

- Statisticians using older software packages tend to not engage in best statistical practices
  - E.g., assume linearity for all predictors
  - Don't routinely incorporate loess, bootstrap, and other methods invented in past 30 years
  - Graphics from hell

Statistical Computing, *continued*

- Systems such as R have rich language for data analysis and graphics
- Avoid point–and–click systems that lead to non-reproducible research
- Supporting the open-source community allows statisticians to give back to the community and to set priorities

Document Management

- LaTeX: the greatest productivity tool
  - Used with text editor; excellent reference, graphics, table, equation methods
  - Dynamically regenerate report when any components change, with cross-references
  - Programmable: conditional text inclusion
- Statistical reporting: marry R and LaTeX: `Sweave` or customized reports

Knowledge Management

- Knowledge is cumulative but constantly updated; memories are imperfect
- `wiki`: collaboration web services (e.g., `twiki.org`)
- Shared authorship/responsibility

Abstract

This talk deals with principles derived from over 30 years of applying statistics to biomedical research, collaborating with clinical and basic biological researchers and epidemiologists. The principles relate to statistical efficiency , bias, validity, robustness, interpretation of statistical results, multivariable predictive modeling, statistical computing, and graphical presentation of information. Topics to be discussed include respecting continuous variables, avoiding non-descriptive statistics, problems associated with filter ing out negative results, overfitting, shrinkage, adjusting P-values for multiple comparisons without adjusting point estimates for same, and the false promise of multi-stage estimation and testing procedures, related to the use of bogus conditional techniques for computing what is advertised as unconditional variances or type I errors.

References

[1] David Brownstone. Regression strategies. In *Proceedings of the 20th Symposium on the Interface between Computer Science and Statistics*, pages 74–79, Washington, DC, 1988. American Statistical Association.

[2] C. Chatfield. Avoiding statistical pitfalls (with discussion). *Statistical Science*, 6:240–268, 1991.

[3] J. J. Faraway. The cost of data analysis. *Journal of Computational and Graphical Statistics*, 1:213–229, 1992.