Biost 518
Applied Biostatistics II

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 11:
Review

March 10, 2006

Purpose of Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)
- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)

First Stage of Scientific Investigation

- Hypothesis generation
  - Observation
    - Measurement of existing populations
    - Disadvantages:
      - Confounding
      - Limited ability to establish cause and effect

Further Stages of Scientific Investigation

- Refinement and confirmation of hypotheses
  - Experiment
    - Intervention
    - Elements of experiment
      - Overall goal
      - Specific aims (hypotheses)
      - Materials and methods
      - Collection of data
      - Analysis
      - Interpretation; Refinement of hypotheses

Statistical Questions

- Clustering of observations
- Clustering of variables
- Quantification of distributions
- Comparing distributions
- Prediction of individual observations

Statistical Tasks

- Understand overall goal
- Refine specific aims (stat hypotheses)
- Materials and methods: Study design
- Collection of data: Advise on QC
- Analysis
  - Describe sample (materials and methods)
  - Analyses to address specific aims
- Interpretation

Statistical Analysis

- Descriptive statistics
  - (Sampling plan)
  - Materials and methods
  - Address scientific question
- Inferential statistics
  - Point estimates
  - Interval estimates (quantify precision)
  - Decision analysis (hypothesis tests)

Purpose of Descriptive Statistics

- Identify errors in measurement, data collection
- Characterize materials and methods
- Assess validity of assumptions needed for analysis
- Straightforward estimates to address scientific question
- Hypothesis generation

Descriptive Methods

- Type of Measurement
  - Binary, nominal, ordered, continuous, censored
- Type of description
  - Univariate
    - Location, spread, skewness, kurtosis
  - Bivariate, trivariate
- Methods
  - Numerical, graphical

Statistical Inference

- Use the sample to make inference about the entire population
  - Inferential estimates
  - Quantify the uncertainty in the estimates computed from the sample
    - To what extent does the random variation inherent in sampling affect our ability to draw conclusions?

Statistical Role

- Experimental results are subject to variability
  - Statistics provides
    - Framework in which to describe general trends
      - Estimates of treatment effect
    - Framework in which to describe our level of confidence in the conclusions drawn from the experiment
      - Measures of the precision of our estimates
    - Estimates of the generalizability of the results

Population Parameters

- Scientific questions are typically answered by making inference about some population parameter, e.g.
  - Mean
  - Geometric mean
  - Median
  - Proportion above threshold
  - Odds above threshold
  - Hazard

Measures of Association

- Most often: difference or ratio of univariate parameters
  - Difference (or ratio) of means
  - Ratio of geometric mean
  - Ratio (or difference) of medians
  - Difference (or ratio) of proportions
  - Odds ratios
  - Hazard ratio (or difference)

Criteria for Summary Measure

- In order of importance
  - Scientifically (clinically) relevant
    - Also reflects current state of knowledge
  - Is likely to vary across levels of the factor of interest
    - Ability to detect variety of changes
  - Statistical precision
    - Only relevant if all other things are equal

Inference

- Generalizations from sample to population
  - Estimation
    - Point estimates
    - Interval estimates
  - Decision analysis (testing)
    - Quantifying strength of evidence
- Frequentist or Bayesian methods exist

Approximate Sampling Distn

- Most often we choose estimators that are asymptotically normally distributed

$$\text{For large } n: \quad \hat{\theta} \dot{\sim} N\left(\text{mean } \theta, \text{var} \frac{V}{n}\right)$$

V is related to average "statistical information" from each observation

Often: V depends on the value of θ

Typical Method for 100(1-α)% CI

- When estimate is approximately normal

$$100(1-\alpha)\% \text{ confidence interval is } (\theta_L, \theta_U)$$
$$\theta_L = \hat{\theta} - z_{1-\alpha/2}\, s\hat{e}(\hat{\theta})$$
$$\theta_U = \hat{\theta} + z_{1-\alpha/2}\, s\hat{e}(\hat{\theta})$$

$$(estimate) \pm (crit\ val) \times (std\ error)$$

Computing P values using Z

Standardized statistic $\quad Z = \dfrac{est - hyp}{std\ err} = \dfrac{\hat{\theta} - \theta_0}{s\hat{e}(\hat{\theta})} \sim N(0,1)$

Stata commands

Lower one-sided P value $\quad \text{norm}\left(\dfrac{\hat{\theta} - \theta_0}{s\hat{e}(\hat{\theta})}\right)$

Upper one-sided P value $\quad 1 - \text{norm}\left(\dfrac{\hat{\theta} - \theta_0}{s\hat{e}(\hat{\theta})}\right)$

Two-sided P value $\quad 2 \times \text{norm}\left(-\text{abs}\left(\dfrac{\hat{\theta} - \theta_0}{s\hat{e}(\hat{\theta})}\right)\right)$

Aside: Comparing Estimates

- Comparisons across strata or studies
  - This is easy, if estimates are independent and approximately normally distributed

For independent $\hat{\theta}_1 \dot{\sim} N(\theta_1, se_1^2)$; $\quad \hat{\theta}_2 \dot{\sim} N(\theta_2, se_2^2)$

$$\hat{\theta}_1 + \hat{\theta}_2 \dot{\sim} N(\theta_1 + \theta_2,\ \ se_1^2 + se_2^2)$$
$$\hat{\theta}_1 - \hat{\theta}_2 \dot{\sim} N(\theta_1 - \theta_2,\ \ se_1^2 + se_2^2)$$
$$\hat{\theta}_1 / \hat{\theta}_2 \dot{\sim} N\left(\frac{\theta_1}{\theta_2},\ \ \frac{1}{\theta_2^2}\left(se_1^2 + \frac{\theta_1^2}{\theta_2^2} se_2^2\right)\right)$$

Aside: Correlated Estimates

- If estimates are correlated and approximately normally distributed

For correlated $\hat{\theta}_1 \dot{\sim} N(\theta_1, se_1^2)$; $\quad \hat{\theta}_2 \dot{\sim} N(\theta_2, se_2^2)$

$$\omega = corr(\hat{\theta}_1, \hat{\theta}_2)$$

$$\hat{\theta}_1 + \hat{\theta}_2 \dot{\sim} N(\theta_1 + \theta_2,\ se_1^2 + se_2^2 + 2\omega\, se_1\, se_2)$$
$$\hat{\theta}_1 - \hat{\theta}_2 \dot{\sim} N(\theta_1 - \theta_2,\ se_1^2 + se_2^2 - 2\omega\, se_1\, se_2)$$

Two Variable Setting

- Many statistical problems consider the association between two variables
  - Response variable
    - (outcome, dependent variable)
  - Grouping variable
    - (predictor, independent variable)

Uses of Regression

- Two major uses of regression
  - Borrow information to address "sparse data" in some groups
    - E.g., 68 and 70 year olds provide information about 69 year olds
    - Question: How far away do you want to go?
  - Provide a statistical "contrast" to compare distribution of response across groups
    - Think of a "slope" as an <u>average</u> comparison of summary measures per unit difference in the grouping variable

Regression Inference

- Estimates
  - Slope: (average) contrasts across groups
  - Fitted values: estimated summary measure in a group
- Standard errors
- Confidence intervals
- P values testing for
  - Intercept of zero (who cares?)
  - Slope of zero (test for linear trend in summary measures)

Regression Models

- According to the parameter compared across groups
  - Means → Linear regression
  - Geom Means → Linear regression on logs
  - Odds → Logistic regression
  - Rates → Poisson regression
  - Hazards → Proportional Hazards regr
  - Quantiles → Parametric survival regr

Regression vs Two Samples

- When used with a binary grouping variable common regression models reduce to the corresponding two variable methods
  - Linear regression with a binary predictor
    - Classical: t test with equal variance
    - Robust SE: t test with unequal variance (approx)
  - Logistic regression with a binary predictor
    - Score test: Chi squared test for association
  - Cox regression with a binary predictor
    - Score test: Logrank test

Maximal Assumptions

- Independence
  - with robust SE: between identified clusters
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model
  - with robust SE: relaxed for associations
- Regression model accurately describes summary measures across groups
- Shape of distribution same in each group

Transformations of Predictors

- We transform predictors to provide more flexible description of complex associations between the response and some scientific measure
  - Threshold effects
  - Exponentially increasing effects
  - U-shaped functions
  - S-shaped functions
  - etc.

Flexible Modeling of Predictors

- We do have methods that can fit a wide variety of curve shapes
  - Dummy variables
    - A step function with tiny steps
  - Polynomials
    - If high degree: allows many patterns of curvature
  - Splines
    - Piecewise linear or piecewise polynomial
  - Fractional polynomial

Choice of Transformation

- Based on the following criteria
  - Scientific issues
    - Scientific question to be addresses
    - Role of predictors in the scientific question
    - Ease of interpretation
  - Statistical issues
    - Accuracy of model
    - Precision of model
      - Parsimony
    - Overfitting of data (inflate type I error)

Scientific Questions

- Most times:
  - Comparing distribution of response across groups defined by predictor of interest
- Very often, other variables also need to be considered because
  - Comparison is different in strata
  - Groups being compared differ in other ways
  - Less variability of response if we control for other variables

Statistical Role

- Covariates other than the POI are included in the model as
  - Effect modifiers
  - Confounders
  - Precision variables

“Model Building”

- Criteria for inclusion of covariates
  - Scientific question
  - Statistical precision
- But sometimes doing data exploration