

Biost 518

Applied Biostatistics II

.....

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

Lecture 15: Review

March 9, 2007

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

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)

2

First Stage of Scientific Investigation.....

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

3

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

4

Statistical Questions

.....

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

5

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

6

Statistical Analysis

.....

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

7

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

8

Descriptive Methods

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

9

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?

10

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

11

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

12

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)

13

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

14

Inference

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

15

Approximate Sampling Distn

- Most often we choose estimators that are asymptotically normally distributed

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

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

Often : V depends on the value of θ

16

Typical Method for 100(1- α)% CI

- When estimate is approximately normal

100(1- α)% confidence interval is (θ_L, θ_U)

$$\theta_L = \hat{\theta} - z_{1-\alpha/2} s\hat{e}(\hat{\theta})$$

$$\theta_U = \hat{\theta} + z_{1-\alpha/2} s\hat{e}(\hat{\theta})$$

$$(\text{estimate}) \pm (\text{crit val}) \times (\text{std error})$$

17

Computing P values using Z

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

Stata commands

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

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

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

18

Aside: Comparing Estimates

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

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

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

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

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

19

Aside: Correlated Estimates

- If estimates are correlated and approximately normally distributed

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

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

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

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

20

Two Variable Setting

.....

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

21

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 average comparison of summary measures per unit difference in the grouping variable

22

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)

23

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

24

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

25

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
 - Necessary for prediction of summary measures
- Shape of distribution same in each group
 - Necessary for prediction (forecast) intervals

26

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.

27

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

28

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)

29

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

30

Statistical Role

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

31

“Model Building”

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

32

Where do YOU
go from here?

.....

33

.....

–“What a long, strange trip it’s
been...”

– “Truckin’”, Grateful Dead

34

.....

–“Well [I’ve] worn out [my] welcome
with random precision...”

– “Shine on You Crazy Diamond”, Pink Floyd

35

.....

Possible Biostat Courses

- Biost 524: Design of Medical Studies
 - The scientific (and some statistical) issues involved in designing clinical trials
 - Biost 517-518 satisfies prerequisites
 - Taught by Tom Fleming in Spring 07

36

Possible Biostat Courses

.....

- Biost 536: Analysis of Categorical Data
 - More experience dealing with categorical response
 - Biost 517-518 satisfies prerequisites
 - Offered in Autumn

37

Possible Biostat Courses

.....

- Biost 537: Analysis of Survival Data
 - More experience dealing with censored time to event data
 - Biost 517-518 satisfies prerequisites
 - Offered in Winter

38

Possible Biostat Courses

.....

- Biost 540: Analysis of Correlated Response Data
 - Dealing with longitudinal data analysis and other clustered observations
 - Biost 517-518 is adequate preparation
 - If you are planning to take Biost 536, 537, and 524, it might be best to take Biost 540 next year
 - Taught by Patrick Heagerty in Spr 07

39