

Biost 517
Applied Biostatistics I
.....
Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 2:
Statistical Classification
of Scientific Questions

October 2, 2009

1

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

Lecture Outline
.....

- Types of Scientific Questions
 - Clustering cases
 - Clustering variables
 - Quantification of distributions
 - Detecting associations
 - Prediction
- Statistical Tasks

2

Science
.....

- Science is about
 - Discovering laws governing the universe
 - Understanding impact / interplay of laws
 - Proving things to people

3

Statistics
.....

- Uses numbers to address scientific questions
 - Describe general tendencies and trends
 - Quantifies our knowledge in scientific laws

4

Scientific Method: Key Elements

- Overall goal
- Specific aims (hypotheses)
- Materials and methods
- Collection of data
- Analysis
- Interpretation

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 Classification of Scientific Questions

.....
 My claim:
 All scientific questions addressed
 with statistics are one of 5 types

7

General Classification

- In order typically used in a new area of science:
 - Clustering of observations
 - Clustering of variables
 - Quantification of distributions
 - Comparing distributions
 - Prediction of individual observations

8

1. Cluster Analysis

- Focus is on identifying similar groups of observations
 - Divide a population into subgroups based on patterns of similar measurements
 - Univariate, multivariate
 - Known or unknown number of clusters
 - (All variables treated symmetrically: No delineation between outcomes and groups)

9

Example: Cluster Analysis

- Potential for different causes for the same clinical syndrome: Glucose in urine
 - Identify patterns of measurements that separate subpopulations of patients with diabetes
 - Age of onset
 - Symptoms at onset (e.g., weight)
 - Auto-antibodies
 - Characteristics of epidemics

10

Example: Cluster Analysis

- Statistical Tasks:
 - Training sample
 - Measure age, change in weight, auto-antibodies, etc.
 - Statistical analysis
 - Cluster analysis
 - Summarize variable distributions within identified clusters
 - (Attach labels?)

11

2. Clustering Variables

- Identifying hidden variables indicating groups that tend to have similar measurements of some outcome
 - Interest in some particular outcome measurement
 - Predictors that imprecisely measure some abstract quality
 - Desire to find patterns in predictors that more precisely reflect the abstract quality

12

Example: Factor Analysis

- Identifying barriers to patient compliance in clinical trials
 - In the Health Behavior Questionnaire, multiple variables might be used to measure
 - Self-perceived health; social support; depression
 - Desire is to
 - Find subset of questions that would suffice
 - Identify hidden variables that affect compliance

13

Example: Factor Analysis

- Statistical Tasks:
 - Training sample
 - Measure response to questionnaire
 - Statistical analysis
 - Factor analysis (principal components)
 - Report contribution to factors, factor loadings
 - (Attach labels?)
 - (Draw conclusions about importance of latent variables?)

14

Example: Genomics/Proteomics

- Combination of clustering cases and variables
 - Measure expression of 10,000 genes on (usually small) number of patients
 - Identify genes that tend to act the same way across patients
 - Pathways?
 - Identify groups of patients that tend to have the same patterns of gene expression
 - Subtypes of disease?

15

3. Quantifying Distributions

- Focus is on distributions of measurements within a population
 - Scientific questions about tendencies for specific measurements within a population
 - Point estimates of summary measures
 - Interval estimates of summary measures
 - Quantifying uncertainty
 - Decisions about hypothesized values

16

Example: Estimate Proportions

- Proportion of women among patients with primary biliary cirrhosis
 - Serious liver disease often leading to liver failure
 - Unknown etiology
 - Characterizing types of people who suffer from disease may provide clues about causes
 - (About 90% of patients with PBC are women)

17

Example: Estimate Proportions

- Statistical Tasks
 - Sample of patients (from registry?)
 - Measure demographics, etc.
 - Statistical analysis
 - Best estimate of the proportion
 - Quantify uncertainty in that estimate
 - Compare to the known proportion of women in the general population (approximately 50%)?

18

Example: Estimation of Median

- Median life expectancy of patients newly diagnosed with stage II breast cancer
 - Want to know prognosis
 - Judging public health risks
 - Patients' planning (?really prediction)

19

Example: Estimation of Median

- Statistical Tasks
 - Sample of patients newly diagnosed with stage II breast cancer
 - Follow for survival time (may be censored)
 - Statistical analysis
 - Best estimate of the median survival (K-M?)
 - Quantify uncertainty in that estimate
 - Compare to some clinically important time range (e.g., 10 years)

20

4. Comparing Distributions

- Comparing distributions of measurements across populations
 - 4a. Identifying groups that have different distributions of some measurement
 - 4b. Quantifying differences in the distribution of some measurement across predefined groups (effects or associations)
 - 4c. Quantifying differences in effects across subgroups (interactions or effect modification)

21

4a. Identifying Groups

- Identifying groups that have different distributions of some measurement
 - Focus is on some particular outcome measurement
 - Identify groups based on other measurements
 - E.g., quantifying distributions within subgroups
 - E.g., stepwise regression models
 - (cf: Cluster analysis where all measurements are treated symmetrically)

22

Example: Identifying Groups

- Chromosomal abnormalities associated with ovarian cancer
 - Cytogenetic analysis of dividing cells identifies regions of the chromosomes with defects
 - Cancer is caused by some defects, and cancer causes other defects
 - Approximately 370 identifiable regions
 - Which of the regions are the most promising to explore in more focused studies?

23

Example: Identifying Groups

- Statistical Tasks:
 - Sample of cancer tissues
 - Measure type of cancer (ovarian, melanoma, etc.)
 - Measure chromosomal defects
 - Statistical analysis
 - Stepwise regression models of chromosomal abnormalities predicting cancer type
 - (Use p values to rank interest in particular regions?)

24

Example: Identifying Groups

- Risk factors for diabetes
 - Variables most associated with diabetes risk may give clues about etiology and eventual prevention

25

Example: Identifying Groups

- Statistical Tasks
 - Sample subjects to measure risk factors and disease prevalence
 - Cohort study
 - Case-control study
 - Statistical analysis
 - Stepwise model building
 - (Rank most interesting variables by p value?)

26

4b. Detecting Associations

- Associations between variables – distributions of one variable differ across groups defined by another
 - Existence of differences
 - Direction of tendency of effect
 - First, second order relationships in a summary measure
 - Characterization of dose-response in a summary measure

27

Definition of an Association

- The distributions of two variables are not independent
 - Independence: Equivalent definitions
 - Probability of outcome and exposure is product of
 - Overall probability of outcome, and
 - Overall probability of exposure
 - Distribution of exposure is EXACTLY the same across ALL outcome categories
 - Distribution of outcome is EXACTLY the same across ALL exposure categories

28

Summary Measures

- Generally we consider some summary measure of the distribution
 - For instance, when we use the mean, we show an association by showing either
 - Mean outcome differs across exposure groups
 - Mean exposure differs across outcome groups

29

Justification

- This works, because if two distributions are the same, ALL summary measures should be the same
 - If some summary measure is different, then we know the distributions are different
- HOWEVER: This means that it is easier to prove an association, than to prove no association

30

Example: Detecting Association

- Effect of blood cholesterol levels on risk of heart attacks
 - Understanding etiology of heart attacks may lead to prevention and/or treatment strategies

31

Example: Detecting Association

- Statistical tasks
 - Measure risk factors, MIs on sample
 - Cohort or case-control sample
 - Statistical analysis
 - Regression model (possibly adjusted)
 - Cohort: Incidence of MIs across cholesterol levels
 - Case-control: Cholesterol levels across MI status
 - (Comparison can be at many levels of detail)
 - Quantify estimates, precision, confidence in decisions

32

4c. Detecting Effect Modification

- Quantifying differences in effects across subgroups (interactions or effect modification)
 - Existence of interaction
 - Direction of interaction (synergy, antagonism)
 - Quantification of exact relationship of interaction

33

Example: Effect Modification

- Identifying whether effect of cholesterol on heart attacks differs by sex
 - Comparing association between blood cholesterol level and incidence of heart attacks between sexes
 - Quantify association in men
 - Quantify association in women
 - Compare measures of association

34

Approach Common to #3 & #4

- In answering each scientific question, statistics typically provides four numbers
 - Best estimate
 - “Best” can be defined by frequentist or Bayesian criteria
 - Interval describing precision
 - Confidence interval or Bayesian credible interval
 - Quantification of belief in some hypothesis
 - P value or Bayesian posterior probability

35

Example: Detecting Association

- Association between sex and prevalence of MI in elderly population
 - 59 of 366 males have had MI: 16.1%
 - 32 of 367 females have had MI: 8.7%
 - Association measured by difference
 - Best estimate: Prevalence 7.4% higher in males
 - Interval estimate: Between 2.7% and 12.2%
 - (95% confidence interval)
 - Strength of evidence: P value = 0.002
 - If there were no real difference, the observed data is pretty unlikely: Probability of this data is 0.002

36

5. Prediction

- Focus is on individual measurements
 - Point prediction:
 - Best single estimate for the measurement that would be obtained on a future individual
 - Continuous measurements
 - Binary measurements (discrimination)
 - Interval prediction:
 - Range of measurements that might reasonably be observed for a future individual

37

Example: Continuous Prediction

- Creatinine clearance
 - Creatinine
 - Breakdown product of creatine
 - Removed by the kidneys by filtration
 - Little secretion, reabsorption
 - Measure of renal function
 - Amount of creatinine cleared by the kidneys in 24 hours

38

Example: Continuous Prediction

- Problem:
 - Need to collect urine output (and blood creatinine) for 24 hours
- Goal:
 - Find blood, urine measures that can be obtained instantly, yet still provide an accurate estimate of a patient's creatinine clearance

39

Example: Continuous Prediction

- Statistical Tasks:
 - Training sample
 - Measure true creatinine clearance
 - Measure sex, age, weight, height, creatinine
 - Statistical analysis
 - Regression model that uses other variables to predict creatinine clearance
 - Quantify accuracy of predictive model
 - (Mean squared error?)

40

Example: Discrimination

.....

- Diagnosis of prostate cancer
 - Use other measurements to predict whether a particular patient might have prostate cancer
 - Demographic: Age, race, (sex)
 - Clinical: Symptoms
 - Biological: Prostate specific antigen (PSA)
 - Goal is a diagnosis for each patient

41

Example: Discrimination

.....

- Statistical Tasks:
 - Training sample
 - “Gold standard” diagnosis
 - Measure age, race, PSA
 - Statistical analysis
 - Regression model that uses other variables to predict prostate cancer diagnosis
 - Quantify accuracy of predictive model
 - (ROC curve analysis?)

42

Example: Interval Prediction

.....

- Determining normal range for PSA
 - Identify the range of PSA values that would be expected in the 95% most typical healthy males
 - Age, race specific values

43

Example: Interval Prediction

.....

- Statistical Tasks:
 - Training sample
 - Measure age, race, PSA
 - Statistical analysis
 - Regression model that uses other variables to define prediction interval
 - (Mean plus/minus 2 SD?)
 - (Confidence interval for quantiles?)
 - Quantify accuracy of predictive model
 - (Coverage probabilities?)

44

Comment About Prediction

.....

- For me to consider a problem to be purely a prediction problem, interest must lie solely in the predicted value, and not in the way that value was obtained
 - E.g., in weather prediction, we might just want to know the weather tomorrow
 - We won't be trying to impress upon our audience the way it should be predicted
 - I do not think this is very often the case

45

Statistical Tasks

.....

46

Statistical Tasks

.....

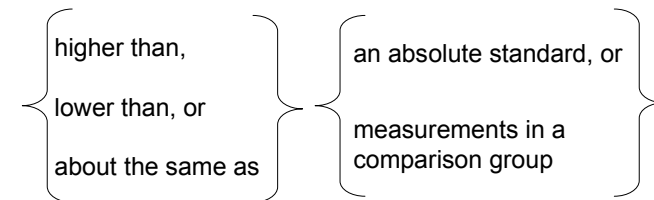
- Statistical considerations come into play in all stages of scientific studies
 - Study Design
 - Data analysis
 - Descriptive statistics
 - Inferential statistics (quantifying precision)
 - Interpretation and reporting of results

47

Scientific Hypotheses

.....

- Usual statement:
 - The intervention when given to the target population will tend to result in outcome measurements that are



48

Refining Scientific Hypotheses

- Statistical hypotheses precisely define
 - the intervention
 - the outcome
 - advise on precision of measurement
 - the target population(s)
 - covariates
 - “tend to” (the standards for comparison)
 - summary measures
 - relevance of absolute or relative standards

49

Study Design: Sampling Plan

- Choosing a method for collecting data
 - Observational vs interventional
 - Cross sectional vs longitudinal
 - Retrospective vs prospective
 - Cohort vs case control
 - Independent vs matched measurements
 - Fixed sample vs sequential
 - Sample size

50

Treatment of Variables

- Measure and compare distribution across groups (response variable in regression)
- Vary systematically (intervention)
- Control at a single level (fixed effects)
- Control at multiple levels (fixed or random effects)
 - Stratified (blocked) randomization
- Measure and adjust (fixed or random effects)
- Treat as “error”

51

Statistical Analysis

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

52