Biost 517
Applied Biostatistics I

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

Lecture 8:
Introduction to Inference

October 21, 2005

1

Lecture Outline

- Statistical Inference
  - Role of Statistical Inference
  - Hierarchy of Experimental Goals
  - Statistical Criteria for Evidence

2

Role of Statistical Inference

3

Statistical Goals of Studies

- Clustering of measurements across variables
- Clustering of variables
- Quantify summary measures of distributions
- Comparison of distributions across groups
  - Interactions
- Prediction of values
  - Single best estimate; interval estimates

4

Use of Samples

- Data is sampled from a population
  - Sampling schemes
    - Observational studies
      - Cross-sectional; cohort; case-control
      - Interventions
  - Time of observation
    - Single point in time
    - Longitudinal

5

Descriptive Statistics

- Purpose of descriptive statistics
  - Detection of errors
  - Materials and methods
  - Validity of methods used in analysis
  - Estimates of association, etc.
  - Hypothesis generation

6

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?

7

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

8

Point Estimates

- Optimal estimates of population summary measures (parameters) or future observations
  - Single best estimate: "Point estimate"
    - Prediction
      - Categorical data: Discrimination, classification
      - Continuous data
    - Population parameters
      - E.g., mean, median, etc.
  - (We must define what we mean by "best")

9

Ex: Estimation of Parameters

- Prognosis in prostate cancer
  - "Parameter" is some summary measure of the population's distribution
    - A descriptive statistic for the entire population
    - E.g., mean, median, proportion above threshold
- Usually use a sample summary measure to estimate the population parameter
  - E.g., Kaplan-Meier estimate of median

10

Ex: Categorical Prediction

- Diagnosis of disease based on laboratory values
  - Type of disease is a categorical variable
  - Use laboratory values to classify patients according to their type of disease
    - (Discriminate between diseases)
  - Obtain training sample in which both type of disease and laboratory values are known
    - Derive a prediction (classification, discrimination) rule

11

Ex: Continuous Prediction

- Creatinine clearance from more easily measured laboratory values
  - Creatinine clearance is a continuous variable
  - Use a single patient's laboratory values to estimate that patient's creatinine clearance
  - Obtain a training sample in which both true creatinine clearance and other laboratory values are known
    - Derive a prediction rule base on mean or median within groups defined by lab values

12

Ex: Prediction Intervals

- Normal range of time delay until arrival of Somatosensory Evoked Potential (SEP)
  - "Normal range" might be defined as the central 95% of the distribution of measurements for a healthy population
  - Goal is to estimate two population parameters
    - 2.5th percentile
    - 97.5th percentile

13

Precision of Estimates

- Choose "best" method for estimation
- Determine how good we were now
  - Quantify confidence/uncertainty in estimates
- Methods will depend upon the type of inference
  - Estimation of population parameters
  - Prediction of individual measurements
    - Categorical
    - Continuous

14

Precision of Parameter Estimates

- Two approaches
  - "Frequentist"
    - What is the variability of the estimate across repeated experiments?
      - Standard error = standard deviation of an estimate
      - Confidence interval = range of values leading to data like this
  - "Bayesian"
    - What is the probability that the true value is in some range?

15

Precision of Continuous Predictions

- "Frequentist"
  - Average absolute error
  - Average squared error
- "Bayesian"
  - Probability of being within a certain tolerance

16

Precision of Classification

- The probability of making an error
  - Overall error rate
    - Proportion of subjects incorrectly classified
    - Depends on frequency of each category
      - Estimated from cross-sectional study?
  - Conditional error rates
    - For each category, proportion of subjects incorrectly classified
      - By disease status (from case-control studies?)
      - By test status (from cohort studies?)

17

Ex: Syphilis and VDRL

- Overall error rate
  - Proportion of subjects incorrectly classified
    - Pr (Pos and Healthy) + Pr (Neg and Syphillis)
- Conditional error based on diagnosis
  - False Positives: Pr (Pos among Healthy)
    - "Specificity" is 1 – False Positive rate
  - False Negatives: Pr (Neg among Diseased)
    - "Sensitivity" is 1 – False Negative rate
- Conditional error based on test result
  - Positive Predictive Value: Pr (Disease among Pos)
  - Negative Predictive Value: Pr (Healthy among Neg)

18

Ex: Cross-sectional Study

- Hypothetical random sample of 1000 STD patients

		Syphilis		
		Yes	No	Tot
VDRL	Pos	270	14	284
	Neg	30	686	716
Total		300	700	1000

19

Ex: Cross-sectional Study

- Valid estimates for inference from cross-sectional study:
  - Prevalence of syphilis (at that clinic): 30.0%
  - Overall error rate: 4.4%
  - Sensitivity: Pr (Pos | Dis) = 270 / 300 = 90.0%
  - Specificity: Pr (Neg | Hlth) = 686 / 700 = 98.0%
  - Pred Val Pos: Pr (Dis | Pos) = 270 / 284 = 95.1%
  - Pred Val Neg: Pr (Hlth | Neg) = 686 / 716 = 95.8%

20

Ex: Sampling by Test Result

- Sample 500 positive subjects and 500 negative subjects at STD clinic (cohort study?)

		Syphilis Yes	No	Tot
VDRL	Pos	475	25	500
	Neg	21	479	500
Total		496	504	1000

21

Ex: Sampling by Test Result

- Valid estimates for inference from study based on sampling according to test result:
  - Prevalence of syphilis (at that clinic): NA
  - Overall error rate: NA
  - Sensitivity: Pr (Pos | Dis) = NA
  - Specificity: Pr (Neg | Hlth) = NA
  - Pred Val Pos: Pr (Dis | Pos) = 475 / 500 = 95.0%
  - Pred Val Neg: Pr (Hlth | Neg) = 479 / 500 = 95.8%

22

Ex: Sampling by Disease Status

- Sample 500 subjects with syphilis and 500 healthy subjects (case-control?)

		Syphilis Yes	No	Tot
VDRL	Pos	450	10	460
	Neg	50	490	540
Total		500	500	1000

23

Ex: Sampling by Disease Status

- Valid estimates for inference from study based on sampling according to disease status:
  - Prevalence of syphilis (at that clinic): NA
  - Overall error rate: NA
  - Sensitivity: Pr (Pos | Dis) = 450 / 500 = 90.0%
  - Specificity: Pr (Neg | Hlth) = 490 / 500 = 98.0%
  - Pred Val Pos: Pr (Dis | Pos) = NA
  - Pred Val Neg: Pr (Hlth | Neg) = NA

24

An Aside: A Generalization

- The previous example had a decision rule based on a binary variable (VDRL)
- With a continuous variable, we usually define a threshold
  - E.g., PSA > 4 for prostate cancer diagnosis
- Sensitivity, specificity will depend on threshold
  - Receiver operating characteristic (ROC) curves consider all possible thresholds

25

Decisions (Hypothesis Testing)

- We often use a statistical analysis to make a binary (yes / no) decision about a hypothesis
  - Precision of our decision is measured by conditional error rates
    - Analogy with categorical prediction

26

Hierarchy of Experimental Goals

27

Ideal: Deterministic Results

- Determine the exact value of a measurement or population parameter
  - Prediction: What will the value of a future observation be?
  - Comparing groups: What is the difference between response across two populations?
- Problem: In the real world, we do not observe the same outcome for all subjects
  - Hidden (unmeasured) variables
  - Inherent randomness

28

2nd Choice: Describe Tendency

- Probability model for response with summary measure for outcomes
  - Phrase scientific question in terms of summary measure
    - Prediction: What is the probability that a future observation will be some value?
    - Within groups: What is the average response within the group?
    - Comparing groups: What is the difference in average response between groups

29

Choice of Summary Measure

- Often we have many choices
  - Example: Treatment of high blood pressure
    - Average
    - Geometric mean
    - Median
    - Percent (or odds) above 160 mm Hg
    - Mean or median time until blood pressure below 140 mm Hg
    - Hazard function

30

Statistical Hypotheses

- Upon choosing a summary measure, the scientific question is stated in terms of the summary measure
  - E.g., Larger mean response might be regarded as “superiority” of a new treatment

31

Criteria for Summary Measure

- Consider (in order of importance)
  - Current state of knowledge about treatment effect
  - Scientific (clinical) relevance of summary measure
  - Plausibility that treatment would affect the summary measure
  - Statistical precision of inference about the summary measure

32

Scientific Importance

- Summary measure for comparison should most often be driven by scientific issues
  - Thresholds may be most important clinically
  - Means allow estimates of total costs/benefits
  - Medians less sensitive to outliers
    - Sometimes clinical importance is not proportional to magnitude of measurements
    - But sometimes, the effect we are trying to detect is greatest on outliers

33

Scientific Importance

- Sometimes choice of summary measure is more arbitrary
  - Types of scientific questions
    - Existence of an effect on the distribution
    - Direction of effect on the distribution
    - Linear approximations to effect on summary measure
    - Quantifying dose-response on summary measure
  - Only last two need dictate a choice of summary measure

34

2nd Choice: Problem

- The distribution (or summary measure) for the outcome is not directly observable
  - Use a sample to estimate the distribution (or summary measure) of outcomes
  - Such an estimate is thus subject to sampling error
    - We want to quantify our uncertainty

35

3rd Choice: Bayesian Methods

- Use the sample to estimate the probability that the hypotheses are true
  - Probability of hypotheses given the observed data
  - Such a Bayesian approach is analogous to the problem of diagnosing disease in patients using a diagnostic procedure

36

Diagnostic Testing

- We most often characterize the sensitivity and specificity of a diagnostic test
  - Sensitivity of test: Positivity in diseased
    - Sample a group of subjects with the disease
    - Estimate the proportion who have a positive test result: Pr (+ | D)
  - Specificity of test: Negativity in healthy
    - Sample a group of healthy subjects
    - Estimate the proportion who have a negative test result: Pr(- | H)

37

Predictive Values

- We are actually interested in the diagnostic utility of the test
  - Predictive value of a positive test: Probability of disease when test is positive
    - Pr (D | +)
  - Predictive value of a negative test: Probability of health when test is negative
    - Pr (H | -)

38

Computing Predictive Values

- Bayes' Rule

$$\Pr(D \mid +) = \frac{\Pr(+ \mid D)\Pr(D)}{\Pr(+ \mid D)\Pr(D) + \Pr(+ \mid H)\Pr(H)}$$

$$\Pr(H \mid -) = \frac{\Pr(- \mid H)\Pr(H)}{\Pr(- \mid H)\Pr(H) + \Pr(- \mid D)\Pr(D)}$$

39

PV+: Relationship to Prevalence

- Need to know sensitivity, specificity, AND prevalence of disease

$$\Pr(D \mid +) = \frac{\Pr(+ \mid D)\Pr(D)}{\Pr(+ \mid D)\Pr(D) + \Pr(+ \mid H)\Pr(H)}$$

$$PVP = \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev)}$$

40

PV-: Relationship to Prevalence

- Need to know sensitivity, specificity, AND prevalence of disease

$$\Pr(H \mid -) = \frac{\Pr(- \mid H)\Pr(H)}{\Pr(- \mid H)\Pr(H) + \Pr(- \mid D)\Pr(D)}$$

$$PVN = \frac{Spec \times (1 - Prev)}{Spec \times (1 - Prev) + (1 - Sens) \times Prev}$$

41

Ex: Syphilis and VDRL

- Typical study: Sample by disease
  - Sensitivity of test: Probability of positive in diseased
    - 90% of subjects with syphilis test positive
    - (Actually depends on duration of infection)
  - Specificity of test: Probability of negative in healthy
    - 98% of subjects without syphilis test negative
    - (Actually depends on age and prevalence of certain other diseases)

42

Ex: PV+, PV- at STD Clinic

- Ex: 1000 patients at STD clinic
  - Prevalence of syphilis 30%
  - PV+: 95% with positive VDRL have syphilis

		Syphilis: Yes	Syphilis: No	Tot
VDRL	Pos	270	14	284
	Neg	30	686	716
Total		300	700	1000

43

Ex: PV+, PV- in Marriage License

- Ex: Screening for marriage license
  - Prevalence of syphilis 2%
  - PV+: 48% with positive VDRL have syphilis

		Syphilis: Yes	Syphilis: No	Tot
VDRL	Pos	18	20	38
	Neg	2	960	962
Total		20	980	1000

44

Bottom Line

- Predictive value of a diagnostic test depends heavily on the prevalence of the disease
  - In typical study (sampling by disease status) we need to use Bayes' Rule to obtain predictive values
    - Prevalence estimated from another study

45

Analogy to Bayesian Inference

- Statistical analysis "diagnoses" an association between variables
  - Association is the true value of parameter
    - Analogous to disease status
  - Estimate association from sample
    - Analogous to the diagnostic test result
  - Compute the probability of hypotheses
    - Analogous to predictive values
    - Need to know prevalence= "prior probability"

46

Implementation

- A generalization of the diagnostic testing situation
  - The estimate of treatment effect is continuous, rather than just positive or negative
  - The parameter measuring a beneficial treatment is continuous, rather than just healthy or diseased
  - Prior distribution is thus an entire distribution (a probability for every possible value of the treatment effect)

47

Issues

- How to choose the prior distribution?
  - As we have seen, the predictive values are very sensitive to the choice of prior distribution
  - Possible remedies:
    - Use data from previous experiments
    - Use subjective opinion or consensus of experts
    - Do a sensitivity analysis over many different choices for the prior distribution
    - Use frequentist approaches

48

4th Choice: Frequentist Methods

- Estimate the behavior of methods over conceptual replications of experiment
  - Calculate the probability of observing data such was obtained in the experiment under the hypotheses
    - Not affected by subjective choice of prior distributions
    - But not really answering the most important question

49

Sampling Distribution

- Frequentist methods consider the "sampling distribution" of statistics over (conceptual) replications of the same study
  - If we were to repeat the study a large number of times (under the exact same conditions) what would be the distribution of the statistics computed from the samples obtained

50

Condition on Hypotheses

- Knowing the true sampling distribution requires knowledge of the parameter
  - We can often guess what would happen under specific hypotheses
  - Frequentists characterize the sampling distribution under specific hypotheses
    - Compare the observed data to what might reasonably have been obtained if that hypothesis were true

51

Bayesian vs Frequentist Poker

- Example: When playing poker, I get 4 full houses in a row
  - Bayesian:
    - Knows the prior probability that I might be a cheater before observing me play
    - Knows the probability that I would get 4 full houses for every level of cheating that I might engage in
    - Computes the posterior probability that I was cheating (probability after observing me play)
    - If that probability is high, calls me a cheater

52

Bayesian vs Frequentist Poker

- Example: When playing poker, I get 4 full houses in a row
  - Frequentist:
    - Hypothetically assumes I am not a cheater
    - Knows the probability that I would get 4 full houses if I were not a cheater
    - If that probability is sufficiently low, calls me a cheater

53

Tradeoffs

- Bayesian: A vague (subjective) answer to the right question
  - How could the Bayesian know my propensity to cheat?
- Frequentist: A precise (objective) answer to the wrong question
  - (The frequentist would give the same answer even if it were impossible that I were a cheater)

54

Tradeoffs

- In fact, there is no real reason to regard tradeoffs as necessary.
- Both approaches contribute complementary information about the strength of statistical evidence.
- It is valid to consider both measures.

55

Bayesian vs Frequentist

- Bayesian inference:
  - How likely are the hypotheses to be true based on the observed data (and a presumed prior distribution)?
- Frequentist inference:
  - Are the data that we observed typical of the hypotheses?

56

Statistical Criteria for Evidence

- At the end of the study analyze the data to provide
  - Estimate of the treatment effect
    - Single best estimate
    - Range of reasonable estimates
  - Decision for or against hypotheses
    - Binary decision
    - Quantification of strength of evidence

57

Point Estimates

- Frequentist methods: using the sampling density for the data
  - Find estimates which minimize bias
    - Difference between true value and average estimate across replicated trials
  - Find estimates with minimal variance
  - Find estimates which minimize mean squared error
- Bayesian methods
  - Use mean, median, or mode of posterior distribution of ? based on some prespecified prior

58

Interval Estimation

- Frequentist confidence intervals
  - Find all values of ? such that it is not unusual to obtain data as extreme as that which was observed
- Bayesian interval estimates
  - Find a range of ? values such that the posterior probability that ? is in that range is high

59

Criteria for Decisions

- Frequentist hypothesis tests
  - Reject hypothesis that $? < ?_0$ if the probability of obtaining the observed data (or more extreme) is low when that hypothesis is true
- Bayesian hypothesis tests
  - Reject hypothesis that $? < ?_0$ if the posterior probability of that hypothesis is low when the observed data is obtained

60

Quantify Evidence for Decision

- Hypothesis testing
  - Based on a statistic T which tends to be large for large θ and an observed value T = t

$$\Pr(T \geq t \mid \theta_0)$$

- Bayesian Methods
  - Based on a presumed prior distribution for θ and the observed observed statistic T = t

$$\Pr(\theta \leq \theta_0 \mid T = t)$$

61

Statistical Criteria for Evidence

- A threshold must be defined for what constitutes a "low" probability
  - Often 5% when considering both too high or too low (a "two-sided" test)
  - Often 2.5% when considering only one direction (a "one-sided" test)

62