Biost 518
Applied Biostatistics II

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

The Use of Statistics to Answer Scientific Questions

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

Biost 518
Applied Biostatistics II

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

Lecture 1:
Course Structure; Overview

January 3, 2007

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

Lecture Outline

- Course Structure
- Overview of Setting
  - Scientific method
  - Case study

Course Overview

Course Structure

- Instructor: Scott S. Emerson, M.D., Ph.D.
- TAs: Mark Giganti, Julian Wolfson
- Time and Place:
  - Lectures: 9:30 - 10:20 am MWF HSB T439
  - Data Analysis:
    - 8:30 - 9:20 am M HSB T478
    - 8:30 - 9:20 am W HSB T478
    - 8:30 - 9:20 am F HSB T359

Textbooks: Optional

- Kleinbaum, et al.: *Applied Regression Analysis*
- Kleinbaum: *Logistic Regression*
- Kleinbaum: *Survival Analysis*

Computer Software

- Extensively used for data analysis
- Students may use any program that will do what is required, however
  - Stata is used heavily in Biostat 536, 537, 540
  - Help will presume the use of Stata
    - I am conversant in S-Plus (very) and SPSS (enough for this class)
  - Other packages may not compute robust standard errors

Stata

- Extremely flexible statistical package
  - Interactive
  - Excellent complement of biostatistical methods
- Graphical, report capabilities suboptimal
- Available in microcomputer lab
- Supplementary info on web page
- Syntax introduced in lectures as needed

Computer Software: Comments

- Designed for people who know statistics, but do not want to write basic functions
  - Tries to be all things to all people
    - Much output that you will not want
    - Much output that I will recommend against

Guiding Principles

- This is a course in biostatistics, not Stata
  - I will tell you how you can get the statistics I teach you to use
    - There are often multiple ways
  - I will not explain every number that appears on the printout

Weekly Homeworks

- Analysis of real data
  - Questions directed toward specific analyses
    - But questions will still be stated in as scientific terms (as opposed to statistical) as possible
  - Work handed in is expected to be organized scientifically
    - I expect nicely formatted tables, figures
    - Unedited Stata output is totally unacceptable

Homework Keys

- Keys to the homeworks will (usually) be available on the web pages
  - Annotated Stata output will typically be included
  - My answers will typically go beyond what I expected you to do
    - You are responsible for any new information that I provide in the homework keys, even if that information is not otherwise presented in class

Discussion Section

- Data Analysis Laboratory
  - Data analysis to answer scientific questions
    - You will be given a scientific question and a data set which was collected to try to answer that question
      - Setting is more realistic than that which is given on written homeworks
    - We will discuss the approach to the whole problem
    - Nothing to hand in, but participation in discussion is expected

Grading

- 25% Homeworks (approx 8)
- 25% One Midterm (in class, closed book)
- 20% Data Analysis and Report
- 30% Final Exam (in class, closed book)

Course Structure

- Biost 517
  - One response variable; one grouping variable
    - One-, two-, K-sample description and inference
    - Simple regression
  - Stratified description and inference
    - Adjustment for confounding, precision
- Biost 518
  - Multivariable regression

Biost 518 Topics

- Review
  - Two variable problem
  - Simple regression
  - Confounding, precision, effect modification
  - Stratified analyses

Biost 518 Topics

- Multiple regression
  - Models, interpretation of parameters
  - Modeling associations
  - Interactions
  - Time varying covariates; clustered data
  - Prediction
  - Missing data
  - Diagnostics
  - Exploratory models

Overview of Setting

Scientific Method

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

Scientific Method: Key Elements

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

Do You Need Statistics?

- Two question test (Both must be YES)
  - In a deterministic world, do YOU know how to answer your question?
    - Is the question answerable in the real world?
    - How do you use a number to answer the scientific question?
  - In a world subject to variation, do YOU know how you would answer your question if you had the entire population?

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 Tasks

Statistical Hypotheses

Statistical Questions

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

Scientific Hypotheses:

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

higher than, lower than, or about the same as	an absolute standard, or measurements in a comparison group

Refining Scientific Hypotheses

- Statistical hypotheses precisely define
  - the intervention (or risk factor)
  - 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

Statistical Role of Variables

- Statistical hypotheses involve
  - "Response" or "Outcome"
    - Can be either the "effect" or the "cause"
  - "Grouping Variable(s)"
    - Primary scientific question
      - Predictor of interest
      - Effect Modifiers
    - Adjustment for covariates
      - Confounders
      - Precision variables

An Aside:

Ability to
Detect Associations

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 the same across all outcome categories
    - Distribution of outcome is the same across all exposure categories

Mathematical Definitions

- Independence: Equivalent definitions
  - Joint probability of outcome O and cause C
    - $Pr(O = o_1, C = c_1) = Pr(O = o_1) \times Pr(C = c_1)$
  - Conditional probability of outcome given cause
    - $Pr(O = o_1 \mid C = c_1) = Pr(O = o_1 \mid C = c_2)$
  - Conditional probability of cause given outcome
    - $Pr(C = c_1 \mid O = o_1) = Pr(C = c_1 \mid O = o_2)$

Establishing Independence

- Consider all events defined by the two variables
  - For each choice of o_1, o_2, c_1, c_2 show either
    - $Pr(O = o_1, C = c_1) = Pr(O = o_1) \times Pr(C = c_1)$,
    - $Pr(O = o_1 \mid C = c_1) = Pr(O = o_1 \mid C = c_2)$, or
    - $Pr(C = c_1 \mid O = o_1) = Pr(C = c_1 \mid O = o_2)$
  - It takes an infinite sample size to prove equality
    - (Relevance to projects from Biost 517)

Detecting Associations

- Instead, we detect associations by showing that two variables are not independent
  - Thus, we show that two distributions are different

Summary Measures

- Generally we consider some summary measure of the distribution
  - E.g., when we use the mean, we show an association by showing either
    - $E(O \times C) \neq E(O) \times E(C)$,
    - $E(O \mid C = c_1) \neq E(O \mid C = c_2)$, or
    - $E(C \mid O = o_1) \neq E(C \mid O = o_2)$

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

Hierarchy of Null Hypotheses

- Strong Null
  - Distribution of response identical in all groups
- Intermediate Null
  - Summary measure identical in all groups
    - Summary measures on a flat line
- Weak Null
  - No linear trend in summary measure across groups
    - On average, summary measures on a flat line

Impact of Study Design

- To establish an association
  - Cohort studies must examine whether
    - $Pr(O \mid C = c_1) \neq Pr(O \mid C = c_2)$
  - Case-control studies must examine whether
    - $Pr(C \mid O = o_1) \neq Pr(C \mid O = o_2)$
  - Cross sectional studies can examine either of the above, as well as whether
    - $Pr(O , C) \neq Pr(O) \times Pr(C)$

Summary Measures

Univariate Summary Measures

- Many times, statistical hypotheses are stated in terms of summary measures for the distribution within groups
  - Means (arithmetic, geometric, harmonic, …)
  - Medians (or other quantiles)
  - Proportion exceeding some threshold
  - Odds of exceeding some threshold
  - Time averaged hazard function (instantaneous risk)
  - …

Comparisons Across Groups

- Comparisons across groups then use differences or ratios
  - Difference / ratio of means (arithmetic, geometric, …)
  - Difference / ratio of proportion exceeding some threshold
  - Difference / ratio of medians (or other quantiles)
  - Ratio of odds of exceeding some threshold
  - Ratio of hazard (averaged across time?)
  - …

Based on Type of Data

- Correspondence to relevance of descriptive statistics
  - Binary or dichotomous:
    - mean (proportion); odds
  - Nominal (unordered categories):
    - frequencies; odds
  - Ordinal (ordered categories):
    - median (quantiles); odds; ? mean
  - Quantitative (addition makes sense):
    - mean; median; proportion > c; hazards, …

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Distribution					
Frequency	OK	OK	OK	OK	
Cum Freq	**boring**		OK	OK	KM
Mode	**boring**	Sample	Sample	Density	
Quantiles	**boring**		OK	OK	KM
Dichotomize Prop / Odds	OK	OK	OK	OK	KM
Means					
Arithmetic	Prop		***	OK	(?KM)
Geometric				OK	
Std Dev	**boring**			OK	
Others				OK	

Joint Summary Measures

- Other times groups are compared using a summary measure for the joint distribution
  - Median difference / ratio of paired observations
  - Probability that a randomly chosen measurement from one population might exceed that from the other
  - …

Commonly Used Parameters

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Entire Distribution	OK	OK	OK	OK	OK
Proportion	OK	Dich	Dich	Dich	Dich
Odds	OK	Dich	Dich, Prop Odds	Dich	Dich
Median			(OK)	OK	OK
Means					
Arithmetic	Prop		(OK)	OK	
Geometric				OK	
Hazard			(OK)	OK	OK
Pr (Y > X)			(OK)	OK	OK

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

Science vs Statistics

- Scientific summary measures
  - Summarize distributions of meaningful measurements
  - Contrasts across populations
    - E.g., a slope
- Statistical measures
  - How precisely we estimate a scientific measure
    - E.g., a P value, correlation

Statistical Tasks

Data Analysis

Descriptive Statistics

- Description of a sample
  - Identification of measurement or data entry errors
  - Characterize materials and methods
  - Validity of analysis methods
    - Assess scientific and statistical assumptions
  - (Straightforward estimates of effects-- inference)
  - Hypothesis generation (inference-- estimation)

Inference

- Generalizations from sample to population
  - Estimation
    - Point estimates
    - Interval estimates
  - Decision analysis (testing)
    - Quantifying strength of evidence

An Aside: Reporting Associations

- Hypothetical study to detect an association between Event B and Exposure F
  - Unexposed: 0 of 5 have Event B
    - » Estimated incidence rate: 0.000
    - » 95% CI for incidence rate: 0.000 – 0.522
  - Exposed: 3 of 5 have Event B
    - » Estimated incidence rate: 0.600
    - » 95% CI for incidence rate: 0.147 – 0.947
  - Fisher's Exact two-sided P: 0.167
  - How would you characterize the presence of an association between these two variables?

WRONG Criteria

- Incorrect criteria for stating the existence of a statistically significant association
  - "Because the confidence intervals overlap, there is no association."
  - (We need to use a P value. The use of confidence intervals in this manner is more complicated.)

Independent CI and Tests

- Rules for **independent** strata
  - IF two independent 95% CI do not overlap
    - THEN we know a statistically significant difference exists (? P less than .006?)
  - IF the 95% CI for one stratum contains the point estimate of the other stratum
    - THEN we know the difference is not statistically significant (? P greater than .16?)
  - OTHERWISE all bets are off
    - Especially: we cannot reverse the above claims

WRONG

- An overstated, purely statistical report
  - "As the P value is greater than 0.05, we conclude that there is no association between exposure F and event B."
  - (We should not conclude that there is no association, because we lacked precision to rule out differences that might be of interest.)

Scientifically USELESS

- A correctly stated, purely statistical report
  - "As the P value is greater than 0.05, we conclude that there is not sufficient evidence to rule out the possibility of no association between exposure F and event B."
  - (Stated correctly, but gives no idea of whether we had ruled out differences that we cared about or we had merely done an abysmal study.)

CORRECT and USEFUL

- Scientific estimates and quantification of statistical evidence
  - "Incidence rates of 60% in the exposed (95% CI: 15% - 95%) and 0% in the unexposed (95% CI: 0% - 52%). Unfortunately, the precision was not adequate to demonstrate that such a large difference in incidence rates would be unlikely in the absence of a true association (P = 0.17)."
  - (These data are not atypical of setting in which F= female and B= giving birth.)

Statistical Tasks

Analysis Methods

Biost 517

- We described tests (and sometimes CI) for comparing parameters across groups
  - Not all are implemented in statistical software, though with a little work they can be obtained in most software packages
  - There are some tests which technically could be applied in certain situations, but it is not very often seen (or recognized)
    - (I have denoted these cases with ?)

Two Independent Samples

		Unordered	Ordered		
	Binary	Nominal	Categ	Quant	Cens
Entire Distn	Chi Sq	Chi Sq	Chi Sq	Kol-Sm	Modif Kol-Sm
Diff in Proportion	Chi Sq	Chi Sq	Chi Sq	Chi Sq	KM
Odds Ratio	Chi Sq; Fish Ex	Chi Sq; Fish Ex	Chi Sq; Fish Ex; Prop Odds	Chi Sq; Fish Ex	KM

Two Independent Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
Diff in Medians			?(Bstrap)	Bstrap	?(Bstrp)
Median Difference			?(Sign)	?(Sign)	
Ratio of Medians					

Two Independent Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
(Diff in) Arithmetic Means (of Diff)	Chi Sq		t test (eq,uneq vrnc)	t test (eq,uneq vrnc)	?(Restr Mean)
(Ratio of) Geometric Means (Ratio)				t test (eq,uneq vrnc) on logs	

Two Independent Samples

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Hazard Ratio				Logrank	Logrank
Pr (Y > X)			Wilcox Rnk Sum	Wilcox Rnk Sum	Modif Wilcox
???			?(Wilcox Sgn Rnk)	?(Wilcox Sgn Rnk)	

Two Matched Samples

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Entire Distn	McNemar (Sign)				
Diff in Proportion	McNemar (Sign)	McNemar (Sign)	McNemar (Sign)	McNemar (Sign)	
Odds Ratio	McNemar (Sign)	McNemar (Sign)	McNemar (Sign)	McNemar (Sign)	

Two Matched Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
Diff in Medians			?(Bstrap)	Bstrap	
Median Difference			Sign	Sign	
Ratio of Medians			?(Bstrap)	Bstrap	

Two Matched Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
(Diff in) Arithmetic Means (of Diff)	McNemar (Sign)		Paired t test	Paired t test	
(Ratio of) Geometric Means (Ratio)				Paired t test on logs	

Two Matched Samples

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Hazard Ratio				Logrank	
Pr (Y > X)			Sign	Sign	
???			Wilcox Sgn Rnk	Wilcox Sgn Rnk	

Regression Methods

- In Biost 518, we extend these methods to the case of the “infinite sample” problem
  - Borrowing information
  - Contrasts across multiple groups

Infinite Samples

- While we don't really ever have (or care) about an infinite number of samples, it is easiest to use models that would allow that in order to handle
  - Continuous predictors of interest
  - Adjustment for other variables

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Entire Distn	Logist				
Diff in Prop	(Linear)	(Linear)	(Linear)	(Linear)	
Odds Ratio	Logist	Logist	Logist; Prop Odds	Logist	

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Diff in Medians					
Median Diffrence					
Ratio of Medians				Param Surv (AFT)	Param Surv (AFT)

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
(Diff in) Arith Means (of Diff)	(Linear)		Linear	Linear	
(Ratio of) Geometric Means (Ratio)				Linear on logs	

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Hazard Ratio				Prop Hazard	Prop Hazard
Pr (Y > X)					
???					

"Everything is Regression"

- The most commonly used two sample tests are special cases of regression
  - Regression with a binary predictor
    - Linear → t test
    - Logistic → chi square (score test)
    - Proportional hazards → logrank (score test)