Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Discussion 1:
Investigating Associations

January 5, 2015

1

Discussion Section

- 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
- Often nothing to hand in, but participation in discussion is required
  - I will often call on students at random
    - It is okay to be wrong, but not okay to be unprepared or inattentive
  - You must inform me if you are attending a different discussion section

2

Reading

- Supplementary materials on class webpages
  - Approach to analyzing a dataset
  - Reporting associations
  - Use of logarithmic transformations

3

Topic for Today

- Today and nearly all quarter:
- What is an association between two variables?
  - Scientifically?
  - Statistically?

4

Statistical Questions

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

5

Statistical Questions

- Clustering of observations
- Clustering of variables
- Quantification of distributions
- **Comparing distributions ➔ Investigating associations**
- Prediction of individual observations

6

Scientific Hypotheses

- Usual statement:

The intervention (exposure) 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 }

7

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

8

Example

- Is serum LDL (the “bad” cholesterol) associated with mortality in older (age $\geq$ 65 years) Americans?
- What are we trying to assess?

9

Example

- Is serum LDL (the “bad” cholesterol) associated with mortality in older (age $\geq$ 65 years) Americans?
- What are we trying to assess?
  - How does the distribution of death differ across groups having different serum LDL?
    - But perhaps similar with respect to other variables to account for confounding or to gain precision
  - How does the distribution of LDL differ across groups having different mortality (early vs late death)?
    - But perhaps similar with respect to other variables to account for confounding or to gain precision

10

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

11

An Aside:

Ability to

Detect Associations

12

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

13

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
    - $\Pr(C = c_1 \mid O = o_1) = \Pr(C = c_1 \mid O = o_2)$

14

Establishing Independence

- Consider all events defined by the two variables
- For every 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
  - Thus “not significant” = “insufficient evidence of to establish an association”
    - not “evidence of no association”

15

Detecting Associations

- Hence, we do not try to show no association
- Instead, we detect associations by showing that two variables are not independent
- Thus, we show that two distributions are different

16

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
  - $E(C \mid O = o_1) \neq E(C \mid O = o_2)$

17

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

18

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

19

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
    - $Pr(O, C) \neq Pr(O) \times Pr(C)$

20

Summary Measures

Example

- How would we statistically detect an association between mortality and serum LDL?

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, …

25

Descriptive Statistics

		Unordered	Ordered		
	Binary	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	

26

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
  - …

27

Commonly Used Parameters

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

28

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

29

Science vs Statistics

- Scientific summary measures
  - Summarize distributions of meaningful measurements
  - Identify populations to compare
    - Differing with respect to some predictor of interest (POI)
      - (By how many units?)
    - Similar with respect to some other variables
      - Confounders, precision variables
  - Contrasts across populations
    - E.g., a slope
- Statistical measures
  - How precisely we estimate a scientific measure
    - E.g., a P value, correlation

30

Statistical Tasks

Data Analysis

31

Descriptive Statistics

- Description of a sample
- Methods will depend on our goal among 5 possible reasons
  - 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)

32

Inference

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

33

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?

34

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.)

35

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

36

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.)

37

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.)

38

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.)

39

Take Home Message

- Ideal: Always give 4 numbers
  - Point estimate
  - Confidence interval (lower, upper bound)
  - P value
- Even when not significant: Give 4 numbers
  - Arguably, it is more important to provide them when not significant
- Be forewarned
  - My job is to improve science into the future
  - If your advisor or the journals you publish in are wrong on this point, I want to make sure you know they are wrong

40

Statistical Tasks

Analysis Methods

41

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 ?)

42

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

43

Two Independent Samples

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

44

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	

45

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)	

46

Two Matched Samples

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

47

Two Matched Samples

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

48

Two Matched Samples

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

49

Two Matched Samples

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

50

Regression Methods

- In Biost 518, we extend these methods to the case of the "infinite sample" problem
- Borrowing information in presence of sparse data
- Contrasts across multiple groups
  - Continuous grouping variables
  - Adjustment for covariates

51

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
  - Compare groups differing in age by 1 year
    - 3 vs 4; 8 vs 9; 11 vs 12 …
  - Figure that comparisons across groups that differ by *k* years will be *k*-fold higher
  - Average all those estimates
- Adjustment for other variables
  - Compare males to females among 30 yo, among 31 yo, …
  - Average all those estimates

52

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	

53

Regression Methods

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

54

Regression Methods

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

55

Regression Methods

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

56

"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)
  - Poisson → two sample test of Poisson rates
  - Proportional hazards → logrank (score test)

57