

Biost 517
Applied Biostatistics I

.....

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

Lecture 3:
Overview of Descriptive Statistics

October 3, 2011

1

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

Where am I going?
.....

- Computing descriptive statistics is generally easy
 - 10th grade WASL

- Understanding what to use when and what they tell you is much harder
 - Important when it comes to inference:
 - “Parameters” are usually descriptive statistics on the population

2

Lecture Outline
.....

- Purpose of Descriptive Statistics
- General Methods
- Types of Measurements

- Types of Summary Measures
 - Univariate
 - Bivariate
 - Three or more variables

3

**Purpose of
Descriptive Statistics**
.....

4

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

5

Purpose #1: Identify errors

.....

- Identify errors in measurement, data collection
- Impossible, improbable, or inappropriate values
 - Univariate: Too low or too high
 - Multivariate: Strange combinations
- Missing data
 - Univariate: Number missing by measurement
 - Multivariate: Predictors of missing data

6

Purpose #2: Materials and Methods

.....

- Characterize materials and methods
- Describe subjects used in study
- Univariate
 - Often broad ranges specified in inclusion/exclusion criteria
 - Want to know exact distributions obtained
- Multivariately
 - Rarely are sample sizes defined for combinations of variables
 - E.g., in the sample are males old and females young?

7

Purpose #3: Validity of Assumptions

.....

- Sampling scheme (of crucial importance)
 - Is our sample relevant to our question?
 - Missing data
 - Confounding
 - Influential cases
- Modeling of dose response (importance depends on question)
 - Effect modification (scientific question or exploration)
 - Linearity of association (estimation in groups or exploration)
- Distributional assumptions (technical and ideally avoided)
 - Within groups
 - e.g., normal, exponential, Poisson distributions
 - Between groups
 - e.g., equal variances, proportional hazards

8

Sampling: Missing Data

.....

- Assumptions about ignorability / nonignorability
 - Ignorable = valid inference can still be obtained without doing anything special
- Mechanisms of missing data
 - MCAR: Missing completely at random (always ignorable)
 - MAR: Missing at random (sometimes ignorable)
 - MNAR: Missing not at random (never ignorable)
- Guiding principle:
 - There is nothing in your data that can tell you about the mechanism or whether it is ignorable
- Approach with descriptive statistics
 - Describe patterns of missingness according to other variables

9

Sampling: Confounding

.....

- Assumptions about presence/absence of confounding
- Confounding: A third variable confuses the estimation of an association between a predictor of interest and the outcome variable
- Definition of a confounder:
 - Associated with the outcome (in a causal manner but not in pathway of interest)
 - Associated with the predictor of interest in the sample

10

Example: Stress and Ulcers

.....

- Alcohol consumption is thought to irritate stomach lining (thus causally associated with outcome)
- Many people drink alcohol when stressed (thus associated with predictor of interest)
- If association truly exists in the population, it may well also exist in the sample
 - But consider randomization which (in some sense) precludes confounding

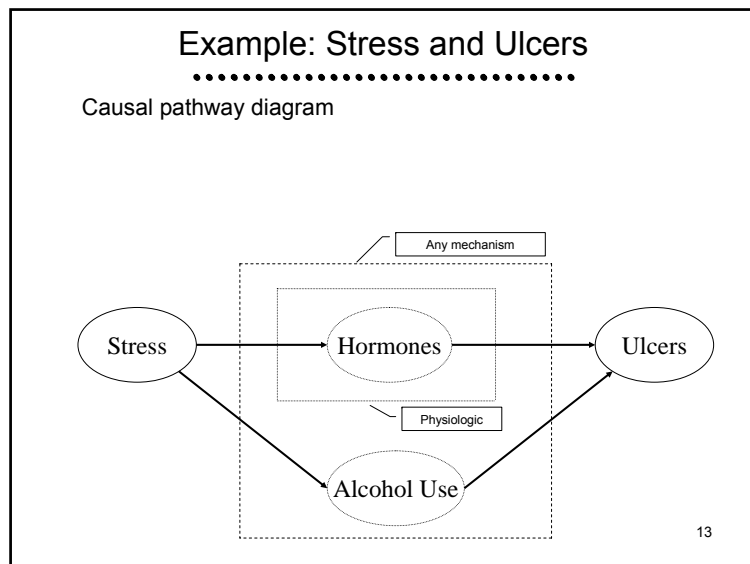
11

Example: Stress and Ulcers

.....

- Is EtOH consumption a confounder?
- In causal pathway of interest?
 - Yes, if interested in all ways stress might cause ulcers
 - And if in causal pathway of interest, we would not want to adjust for EtOH as a confounder
 - No, if only interested in determining whether the physiologic consequences of stress cause ulcers
 - And if not in causal pathway of interest, then EtOH consumption would be confounding our ability to assess physiologic consequences

12



- ### Sampling: Influential Cases
-
- Sometimes our inference about an association is unduly affected by a single observation
 - Outliers: Unusual outcome for the group that they are in
 - E.g., a 6 foot tall three year old
 - Highly leveraged points: Unusual for the sampling scheme
 - E.g., a sample that contains 500 children and one 80 year old
 - Influence: combination of leverage and outliers or nonlinearity
 - If the case is deleted, would cause very different inference
 - Changing the estimated association
 - High leverage and outlier or nonlinearity
 - Changing the statistical significance
 - High leverage or outlier
- 14

- ### Purpose #4: Preliminary Estimates
-
- Estimates for statistical inference
 - Many estimates used in statistical inference are based on sample descriptive statistics
 - E.g., “method of moments” estimators are defined by using sample moments (means, variances, etc.) to estimate population moments
- 15

- ### Purpose #5: Generate Hypotheses
-
- Exploring unanticipated effects
 - Characterization of dose-response
 - Linear
 - U-shaped
 - Threshold
 - Exploring difference in effects across subgroups
 - E.g., is association between treatment and clinical outcome similar in men and women
- 16

General Methods

.....

17

General Process

.....

- From study protocol
 - Describe sampling methods
 - Identify variables
 - Scientific role, statistical role, type of measurement
- From data
 - Univariate statistics
 - Bivariate statistics
 - Three or more variables

18

1. Identify Sampling Scheme

.....

- Describe sampling methods
 - From protocol, rather than data
- Source of data
 - Location(s)
 - Time(s)
 - Selection criteria
 - Screening criteria
 - Inclusion criteria
 - Exclusion criteria

19

Constrained Sample Sizes

.....

- Sample sizes specified by design
 - Overall and/or within prespecified strata
 - E.g., cohort or case-control designs
- Sample sizes reflecting random process
 - Sometimes sampling scheme specifies time and location of sampling, not sample size
 - Allows estimate of prevalence or incidence
 - E.g., sample all cases of new lung cancer in Seattle during 1999
 - Sample size can be used to estimate incidence of lung cancer

20

Common Study Designs

.....

- Cross-sectional studies (surveys)
- Cohort studies
- Case-control studies
- Interventional studies

21

Cross-sectional Studies

.....

- Surveys of subjects sampled from a population
- Real or event time
 - “Real time” = “calendar time”
 - “Event time” = when some event happens
 - birth, marriage, diagnosis, treatment, death
- Efficient for examining
 - Common outcomes and risk factors
 - Associations (not cause and effect)
 - Can estimate prevalence of risk factors and outcomes
 - Overall and within groups

22

Cohort Studies

.....

- Groups defined by risk factor
 - Identified prospectively or retrospectively
- Followed longitudinally for outcome(s)
 - Prospectively into the future, or
 - Retrospectively since some defining event
 - e.g., since being born in a particular hospital in a particular year
- Efficient for examining
 - Common outcomes
 - Many different outcomes for same exposure
 - Associations (not cause and effect)
 - Estimate incidence within risk factor groups
 - Cannot estimate prevalence of risk factor

23

Case-Control Studies

.....

- Groups defined by some outcome event
 - E.g., death, diagnosis of disease
- Characterize prior exposures
 - Longitudinal study into the past
 - How to handle time in exposure?
 - ever / never exposed, cumulative exposure, time since exposure
- Efficient for examining
 - Rare outcomes
 - Many different risk factors for same outcome
 - Associations (not cause and effect)
 - Estimate prevalence of exposure by disease
 - Cannot estimate prevalence of disease

24

Interventional Studies

.....

- Subjects assigned to some intervention
 - Ideally controlled, randomized
- Followed longitudinally for some outcome
 - So a special case of a cohort study
- Efficient for examining
 - Common outcomes
 - Cause and effect

25

Conspiracies Against Laity I

.....

- Random variables
- A “random variable” is some measurement that might vary across subjects
 - Commonly denoted by capital letter or a mnemonic
 - e.g., A or AGE or Age
- We commonly denote the values a random variable can be using lower case letters
 - We talk about “events”, e.g.,
 - $A = a_1$
 - $A \leq a_5$
 - $a_{10} \leq A \leq a_{15}$

26

Conspiracies Against Laity II

.....

- Probability distributions
- We know everything there is to know about a random variable when we can describe the probability of every possible event
- Two common methods
 - Cumulative distribution function (cdf)
 - Know $Pr(Y \leq y)$ for every possible value of y
 - Probability mass function (pmf)
 - Know $Pr(Y = y)$ for every possible value of y

27

Conspiracies Against Laity III

.....

- Conditional probability
 - (The way most people lie with statistics)
- Often we want to talk about the distribution of a random variable within a restricted group
 - E.g., the distribution of weight among males
- Notation: $Wgt | Male$
 - Conditional CDF: $Pr(Wgt \leq w | Male = m)$
 - Conditional *distn* of weight among males
 - $Pr(Wgt \leq w | Male = 1)$
 - Conditional *distn* of weight among females
 - $Pr(Wgt \leq w | Male = 0)$
 - Very different from $Male | Wgt$
 - Probability of being male for a given weight

28

Conspiracies Against Laity IV

.....

- Summary measures

- We often summarize aspects of distributions
 - Mean (or expectation): $E(Y)$
 - Median: $Mdn(Y)$
 - Variance: $Var(Y)$

- Summarizing conditional distributions
 - Conditional mean (or expectation): $E(Y | X=x)$
 - Median: $Mdn(Y | X=x)$
 - Variance: $Var(Y | X=x)$

29

Detecting Associations

.....

- Consider random variables
 - D be the disease state with values (d_1, d_2, \dots)
 - R be a risk factor with values (r_1, r_2, \dots)

- We consider the “statistical questions” that can be answered by study designs
 - Cross-sectional
 - Cohort
 - Case-control

30

Detecting Associations

.....

- Cross-sectional surveys show
 - $E(D | R = r_1) \neq E(D | R = r_2)$, OR
 - $E(R | D = d_1) \neq E(R | D = d_2)$

- Cohort studies sample within risk factors so only consider
 - $E(D | R = r_1) \neq E(D | R = r_2)$

- Case-control studies sample within disease so only consider
 - $E(R | D = d_1) \neq E(R | D = d_2)$

31

Detecting Cause and Effect

.....

- Demonstrated rigorously only through randomized studies

- A characteristic of study design

- There is nothing in the data that can distinguish between randomized studies and observational studies

32

2. Identify Variables of Interest

.....

- Identify variables of interest according to
 - Scientific meaning
 - Statistical role
 - Type of measurement

33

Scientific Meaning of Variables

.....

- Demographic variables
- Measures of exposure
- Measures of concurrent disease
- Measures of severity of disease
 - Cardiovascular function
 - Liver function
 - etc.
- Measures of clinical outcomes
- etc.

34

Statistical Role of Variables

.....

- Outcome (response) variable(s)
 - Primary and surrogates
- Predictor(s) of interest (define main groups)
- Subgroups of interest for effect modification
- Potential confounders
- Variables that add precision to analysis
 - Known to be associated with response
 - Often these are potential confounders
 - may be associated with predictor(s) of interest in sample
- Irrelevant to current question

35

3. Identify Type of Measurement

.....

- The way in which a variable is measured will affect the descriptive statistics that are of interest
 - Binary (dichotomous, Bernoulli)
 - Nominal (unordered categorical)
 - Ordered categorical
 - Quantitative
 - Discrete, interval continuous, ratio continuous
 - Censored

36

Types of Measurements

.....

37

Characterizing Measurements

.....

- Number of possible values
 - One, two, finite, countably / uncountably infinite
- Comparisons between values
 - Unordered, partially ordered, totally ordered
 - Scientific relevance of differences, ratios
- Completeness of measurement
 - Censoring

38

Binary Measurements

.....

- Only two possible values, which can be either
 - Labels, e.g., “Male” or “Female”
 - Coded as numbers, e.g., 1 or 2
- Most often it is statistically advantageous to represent as “indicator variables”
 - Possible values 0 or 1
 - 1 indicates the quality named by the variable
 - E.g., MALE is 1 for males, 0 for females
 - E.g., MARRIED is 1 for married, 0 for single, divorced, widowed, everything else

39

Properties of Binary Measures

.....

- Ordered
 - Differences (but not ratios) of measurements have a scientific interpretation
- The mean of an indicator variable is the proportion of subjects having the corresponding quality
 - Differences of means are scientifically relevant
 - Ratios of means are scientifically relevant
 - (Both differences and ratios of means may have limited ranges of interest for a specific problem)

40

Categorical Measurements

.....

- A finite number of possible values denoting qualities
 - E.g., occupation is laborer, clerical, professional, retired
 - E.g., marital status is single, cohabiting, married, divorced, separated, widowed
 - E.g., stage of cancer is I, II, III, or IV

41

Unordered Categorical

.....

- Unordered: no clear ordering of values can be prespecified
 - E.g., marital status
 - E.g., occupation status (unless used as a surrogate for physical exertion, sun exposure, etc.)

42

Totally Ordered Categorical

.....

- Totally ordered: categories can be qualitatively, but not quantitatively, ordered
 - Neither differences nor ratios have consistent scientific meaning
 - E.g., stage of cancer, degree of swelling

43

Partially Ordered Categorical

.....

- Partially ordered
 - Some categories have clear ordering, but others cannot be
 - E.g., Atypia on Pap smear often has "indeterminate" results
 - E.g., Severity of cancer might involve both grade and stage
 - May be hard to decide which is more severe:
 - Low grade and high stage, or
 - High grade and low stage

44

Means of Categorical Variables

.....

- Descriptively of less interest even for ordered
 - Spacing between categories is not well-defined

- However,
 - Means sometimes can still be used to identify (but not quantify) differences between distributions of categorical variables
 - Means may be particularly attractive in detecting shifts toward higher levels across groups with totally ordered categorical variables

45

Quantitative Variables

.....

- Values represent a (reasonably) precise quantification of some scientific measure

- Values can be
 - Discrete levels
 - No possible measurements between adjacent levels
 - E.g., counts of events
 - Continuous levels
 - E.g., weight
 - Distinction is often more a question of number of levels:
 - Money is measured to nearest \$0.01
 - But often regarded as continuous

46

Interval vs Ratio Measurements

.....

- Generally, differences make sense for all quantitative variables

- Ratios only make sense if measurements are made relative to an absolute zero
 - Age, height, weight have absolute zeroes
 - Temperature has different zeroes in Fahrenheit and Celsius

- Categories of quantitative variables:
 - Interval: Only differences make clear sense
 - Ratio: Both differences and ratios of interest

47

General Use of Ratios

.....

- Ratios have no scientific relevance with interval measurements

- Thus not of great interest descriptively

- May still be of use in identifying differences in distributions across groups
 - E.g., A ratio of temperatures different from 1 indicates different distributions

- Quantifying differences in distributions will be specific to units used
 - Twice as hot in Fahrenheit vs twice as hot in Celsius

48

Censored Variables

.....

- A special type of missing data commonly arises in applications due to censored measurements (the exact value is not always known)
- Right censoring: for some observations it is only known that the true value exceeds some threshold
- Left censoring: for some observations it is only known that the true value is below some threshold
- Interval censoring: for some observations it is only known that the true value is between two thresholds
- (N.B.: Do not use the word “censoring” to mean every type of missing data. It is used to mean partially incomplete data)

49

Example: Right Censoring

.....

- Clinical trial detecting effect of aspirin on cardiovascular death
 - At the time of data analysis, death times have been observed for some subjects
 - At the time of data analysis, some subjects are still alive
- Representation of data using two variables
 - A variable measuring observation time until death or time of analysis, whichever comes first
 - An indicator variable telling which times are death times

50

Types of Summary Measures

.....

- By feature of distribution
 - Typical value (location)
 - Spread of distribution (variability)
 - Symmetry of distribution (skewness)
 - Tendency to extreme values (kurtosis)
 - Depiction of entire distribution
- By number of variables described
 - Univariate
 - Bivariate
 - Higher dimensional

51

52

Univariate Location

- Measures of location ("Typical value")
- Numeric
 - Mode
 - Mean (arithmetic, geometric, harmonic)
 - Median (other percentiles)
 - Proportion exceeding a threshold
 - Odds of exceeding a threshold
- Graphical
 - Mode of density

53

Univariate Spread

- Measures of spread
- Numeric
 - Range (min, max)
 - Interquartile range (25%ile, 75%ile)
 - Variance
 - Standard deviation
- Graphical
 - Box plot
 - Histogram
 - Density

54

Univariate Symmetry

- Measures of symmetry
- Numeric
 - Coefficient of skewness
 - (Compare mean and median, etc.)
- Graphical
 - Histogram
 - Density
 - Box plot

55

Univariate "Heavy Tails"

- Measures of tendency to extreme values
- Numeric
 - Coefficient of kurtosis

56

Univariate Entire Distribution

.....

- Numeric
 - Frequency tables
 - CDF tables
- Graphical
 - Histogram (stem-leaf)
 - Ogive
 - Density estimates
 - Empirical CDF, survival curves
 - Hazards
 - Box plots

57

Univariate Descriptive Statistics

		Binary	Unordered	Ordered		
			Nominal	Categ	Quant	Cens
Entire Distribution	Frequency	OK	OK	OK	OK	
	Cum Freq	boring		OK	OK	KM
	Mode	boring	Sample	Sample	Density	
	Min / Max	boring		boring	OK	
Dichotomize	Proportion (or Odds)	OK	OK	OK	OK	KM
	Quantiles (25 th , Mdn, 75 th)	boring		OK	OK	KM
Means	Arithmetic	(Prop)		***	OK	(?KM)
	Geometric				OK	(?KM)
	Harmonic				OK	(?KM)
	Std Dev	boring			OK	(?KM)
	Skew, Kurt	boring			OK	(?KM)

Bivariate Summary Measures

.....

- Measures of association
- Numeric
 - Stratified univariate descriptives
 - Slope of best fitting line
 - Correlation
 - Rank correlation
- Graphical
 - Least squares line
 - Scatterplot smoother
 - Stratified box plots

59

Bivariate Outliers

.....

- Outliers: Data points far from any others
- Numeric
 - Hat matrix
- Graphical
 - Scatterplot

60

Bivariate Entire Distribution

- Characterization of entire distribution
- Numeric
 - Cross tabulation
- Graphical
 - Scatterplot

61

Three or More Variables

- Measures of association
- Numeric
 - Stratified univariate descriptives
- Graphical
 - Stratified least squares
 - Stratified scatterplot smoothers

62

Three or More Variables

- Measures of interaction (effect modification)
- Numeric
 - Stratified descriptives of bivariate association
- Graphical
 - Stratified least squares
 - Stratified scatterplot smoothers

63

Three or More Variables

- Measures of outlying values
- Numeric
 - Hat matrix

64

Three or More Variables

- Characterization of entire distribution
- Numeric
 - Cross tabulation
- Graphical
 - Stratified scatterplots

65

What Do I Really Use?

- Univariate
 - Number of Missing
 - Mean
 - Standard Deviation
 - Min, Max
 - 25th, 50th (median), 75th percentile
- Bivariate (and Trivariate)
 - Scatterplots (and smooths)
 - Stratified statistics

66