

Biost 517

Applied Biostatistics I

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

Lecture 4: Descriptive Measures of Location

October 5, 2007

1

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

Lecture Outline

-
- Notation
 - Univariate Measures of Location
 - Mode
 - Means (Arithmetic, Geometric, Harmonic)
 - Median (Other Quantiles)
 - Proportions
 - Odds

2

Notation

.....

3

Random Variables

-
- Variables that take on particular values according to a probability distribution
 - The value of the random variables may be a label (sometimes a numeral is used as a label) or a number
 - Mathematicians tend to use X, Y, Z
 - With no scientific problem in mind, I will too
 - For a particular scientific problem, I will tend to use mnemonics, e.g., AGE, HEIGHT, WEIGHT

4

Sample of Measurements

.....

- Measurements made on n subjects
 - Subscripts on a random variable denote the measurements made on different subjects
 - X_1 will be the measurement of X on subject 1
 - X_5 will be the measurement of X on subject 5
 - X_i will be the measurement of X on subject i
 - As a general rule, it is arbitrary which subject is the first, second, etc. so long as we are consistent

5

Random Variables vs Observed

.....

- The notation can seem a little confusing:
 - Random variable $Chol$ might denote the idea of making a measurement of cholesterol
 - Random variable $Chol_i$ might denote the idea of making a measurement on the i -th subject
 - Constant c might denote a particular value of the random variable $Chol$
 - c_i might denote the particular value observed for the measurement on the i -th subject

6

Notation for Ordered Measurements

.....

- Some descriptive statistics make use of the order statistics for a sample:

Data $\{X_1, X_2, \dots, X_n\}$

Order stats $\{X_{(1)}, X_{(2)}, \dots, X_{(n)}\}$

$(X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)})$ are the data in order)

7

Univariate Measures of Location

.....

8

Measures of Central Tendency

.....

- In order (more or less) of usefulness
 - Mean (average; arithmetic mean)
 - Geometric mean
 - Median
 - Proportion exceeding a specific threshold
 - Odds of exceeding a specific threshold
 - Mode

9

“You better think (think),
think about what
you're trying to do...”

Aretha Franklin, “Think”

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

11

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

12

(Arithmetic) Mean

- Definition of sample arithmetic mean:
 - Sum of measurements divided by the number of measurements (average)
 - Notation: Usually denoted by a bar over the variable

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{(X_1 + X_2 + \dots + X_n)}{n}$$

13

Mean: Example

- Data: {1, 3, 6, 3, 2, 3, 6, 7, 1, 1}
 - Number of observations: 10
 - Sum of observations: 33
 - Arithmetic mean: 3.3

14

Mean: Interpretation

- The “point of balance” for the distribution
 - Center of gravity
- If all measurements were the same, the arithmetic mean is the value that they would all be in order to have the same total
- Allows prediction of the total of an arbitrary number of observations
 - Applications in gambling: How much would I expect to win/lose if I play 1,000 times
 - Similarly useful in health care cost analysis

15

Mean: Types of Variables

- Types of variables
 - Defined only for variables that take on numeric values (sum must make sense)
 - Most sensible when differences have scientific interpretation on a constant scale
 - Treats observation of {1,5} similar to {3,3}
 - (5 - 4) should be similar equivalent to (3 - 2), etc.
 - (But see comments regarding comparisons of ordered categorical variables)

16

Mean: Censored Variables

- Not of interest with variables measuring censored times to an event
 - Must know all relevant values exactly in order to compute sum
 - The observation time is a mixture of times to event and times to censoring
 - The indicator of events is measured over varying time periods
 - (Alternative methods using Kaplan-Meier estimate of survivor function (see later) sometimes do allow estimation of mean with censored data)

17

Mean: Descriptive Uses

- By purpose of descriptive statistics
 - Characterizing sample
 - Often used as a “typical value”
 - Note that mean is heavily influenced by large outliers, so sometimes mean does not reflect the quantity desired
 - Assessing validity of assumptions
 - Often we model the mean by assuming linear, quadratic, etc. relationship
 - Most often best measure of potential confounding

18

Mean: Scientific Uses

- Prediction of an individual observation
 - (minimizes squared error loss)
- Clustering:
 - Some algorithms minimize distance to means
- Quantifying distributions:
 - Sometimes best related to the scientific question
 - E.g.: Mean blood pressure and chronic vascular disease
 - E.g.: Studying total health care costs
- Comparing distributions:
 - Sensitive to a wide variety of differences in distr

19

Mean: Ordered Categorical Data

- The mean by itself is not scientifically interpretable
- The mean can still detect differences in the distributions
 - Sensitive to certain tendencies for higher measurements in one group

20

Geometric Mean

- Definition of geometric mean:
 - Exponentiated arithmetic mean of log transformed data
 - (exponentiation and logarithm should be same base)

$$\exp\left(\frac{1}{n} \sum_{i=1}^n \log(X_i)\right) = \sqrt[n]{\prod_{i=1}^n X_i}$$

21

Geom Mean: Descriptive Uses

- The geometric mean is related to the mean of log transformed data
 - All comments about the arithmetic mean need to apply to the log transformed data
- When variables are measured on an exponential scale, geometric means tend to be more stable
 - Serial dilutions used for measuring titers
- Ratios of geometric means tend to be more stable than ratios of arithmetic means.
 - The log of a ratio is the difference of logs

22

Geom Mean: Scientific Uses

- Sometimes a better measure of “typical” values for skewed data (with large outliers)
- Greater statistical precision than mean when standard deviations are proportional to mean

23

Harmonic Mean

- Definition of harmonic mean
 - Reciprocal of the arithmetic mean of the reciprocals of the data

$$\frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{X_i}}$$

24

Harmonic Means: Uses

- The harmonic mean is related to the mean of the reciprocal of the data
 - All comments about the arithmetic mean need to apply to the reciprocal of the data
- The harmonic mean sometimes has scientific interpretation
 - E.g., in electricity, resistance of parallel resistors
 - E.g., in studying vascular flow and blood pressure

25

Median

- Definition:
 - The value that is larger than half the population and smaller than half the population

$mdn(X)$ is any value M such that

$$\Pr(X \leq M) \geq 0.5$$

$$\Pr(X \geq M) \geq 0.5$$

26

Calculation of Sample Median

- In order to find a unique sample median, we usually use the following definition for uncensored data

$$n \text{ odd} : mdn(X) = X_{\left(\frac{n+1}{2}\right)}$$

$$n \text{ even} : mdn(X) = \frac{1}{2} \left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)} \right)$$

27

Median: Examples

- Example (odd number of observations)

- Data: {1, 3, 6, 3, 2, 3, 6, 7, 1, 1, 5}
- Order statistics: {1, 1, 1, 2, 3, 3, 3, 5, 6, 6, 7}
- Median: 3

- Example (even number of observations)

- Data: {1, 6, 2, 3, 6, 7, 1, 5}
- Order statistics: {1, 1, 2, 3, 5, 6, 6, 7}
- Median: 4

28

Median: Type of Variables

- Concept defined for any ordered variable
- Special methods (Kaplan-Meier estimates) must be used with censored data
 - The sample median is not of interest
 - The observation time is a mixture of times to event and times to censoring
 - The indicator of events is measured over varying time periods
 - More often able to estimate median from censored data than mean

29

Median: Descriptive Uses

- Characterizing distribution of sample
 - “Typical” value, especially when number of observations above or below is most meaningful scientifically
 - Less influenced (not influenced) by outliers

30

Median: Scientific Uses

- Prediction: If median is “better behaved”
 - (tends to minimize absolute error loss)
- Quantifying distributions:
 - Not sensitive to “outliers”
 - Estimated more efficiently in presence of outliers
- Comparing distributions:
 - Not sensitive to “outliers”
 - Sometimes more precision for skewed data
 - Not as useful when distn only differ in the tails

31

Proportions

- Dichotomize observations by defining a group of interest
 - Ordered: According to whether they exceed some threshold
 - E.g., Age > 50
 - Unordered: Divide into groups
 - E.g., Marital status separated or divorced

$$Y_i = \begin{cases} 1 & \text{if in group of interest} \\ 0 & \text{otherwise} \end{cases}$$

32

Proportions

- Proportion of subjects in the group of interest is the mean of the dichotomized data

$$\hat{p} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

33

Proportion: Type of variables

- Types of variables
 - Can be defined for nearly any variable
 - Dichotomization must be scientifically interpretable
 - Not of interest with variables measuring censored times to an event
 - The observation time is a mixture of times to event and times to censoring
 - The indicator of events is measured over varying time periods
 - Natural for binary variables

34

Proportion: Descriptive Uses

- By purpose of descriptive statistics
 - Useful in characterizing distribution
 - In presence of scientifically meaningful threshold
 - E.g., proportion of subjects with cholesterol below 200
 - When mode is minimum or maximum
 - E.g., proportion of nonsmokers

35

Proportion: Scientific Uses

- Scientific questions
 - Prediction: Discrimination, classification
 - Discrimination
 - Quantifying distributions:
 - Binary variables
 - Scientifically important thresholds, categories
 - Comparing distributions:
 - Binary variables
 - Scientifically important thresholds, categories

36

Odds

- Odds of being in group of interest:

- Definition

- Dichotomize observations as with proportion
- Ratio of proportion to 1 minus proportion

$$\hat{p} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$\hat{o} = \frac{\hat{p}}{1 - \hat{p}}$$

37

Odds: Uses

- As with proportions, except

- Odds is less easily understood by a lay person
 - Odds of rolling 1 on a die: 1/5
- However, odds (actually log odds) is often more convenient in modeling due to greater range of possible values
 - proportion is between 0 and 1
 - odds is between 0 and infinity
 - log odds is any real number

38

Mode

- Definition:

- Discrete data: Most frequently occurring value
- Continuous data: (Local) maximum of density
 - Determined from histogram or density

39

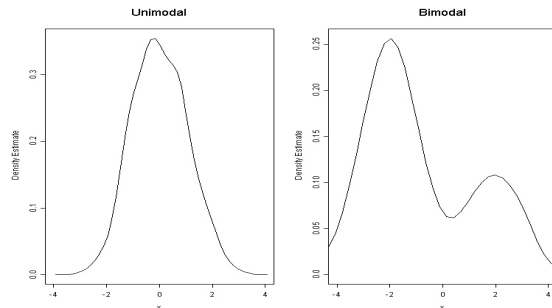
Mode: Discrete Data Examples

- Data: {1, 3, 7, 3, 2, 3, 6, 7}
 - Order statistics: {1, 2, 3, 3, 3, 6, 7, 7}
 - Mode: 3
- Data: {1, 3, 6, 3, 2, 3, 6, 7, 1, 1}
 - Order statistics: {1, 1, 1, 2, 3, 3, 3, 6, 6, 7}
 - Modes: 1, 3
- Data: {1, 0, 9, -3, 2, 3, 6, 7, 8, 4}
 - Order statistics: {-3, 0, 1, 2, 3, 4, 6, 7, 8, 9}
 - Modes: none (or all?)

40

Mode: Continuous Data Example

- Density estimates from samples



41

Mode: Types of Variables

- Types of variables
 - Defined for both categorical and continuous
 - (though definitions differ)
 - Not dependent on ability to order values
 - Not of interest with variables measuring censored times to an event
 - The observation time is a mixture of times to event and times to censoring
 - The indicator of events is measured over varying time periods

42

Mode: Uses

- By purpose of descriptive statistics
 - Sometimes useful in characterizing sample
 - Suitable as a “typical” value, providing it has high enough frequency or density
 - Hypothesis generation
 - Multimodal distributions might indicate mixture of populations
 - Subgroups of subjects might behave differently

43

Mode: Scientific Uses

- Prediction:
 - If mode represents overwhelming majority of distribution
- Cluster analysis:
 - Multimodal distributions often considered mixture of populations
- Quantifying or comparing distributions:
 - Rarely used due to difficult inference

44

Stata Commands

.....

45

Stata: describe

.....

- “describe *varlist*”
 - Provides type of variable
 - Value labels

46

Stata: inspect

.....

- “inspect *varlist*”
 - Line printer histogram
 - Counts of missing vs nonmissing
 - Counts of integer vs noninteger
 - Counts of negative, zero, positive values

47

Stata: summarize

.....

- “summarize *varlist*, format”
 - Table format with number of nonmissing observations, mean, standard deviation, minimum maximum
- “summarize *varlist*, detail format”
 - Additionally provides quantiles, skewness, kurtosis, but not in useful table

48

Stata: means

- “means *varlist*”
 - Table format with arithmetic, geometric, and harmonic means
 - Also gives
 - number of nonmissing observations
 - confidence intervals for inference)
- (Could also obtain geometric mean by log transforming data, taking mean, exponentiating)

49

Stata: centile

- “centile *varlist, cen(20 25 50)*”
 - Table format with requested quantiles
 - Also gives
 - confidence intervals for inference
 - 25th, 75th percentiles may not agree with other Stata functions

50

Stata: tabstat

- “tabstat *varlist, stat(n mean sd min p25 med p75 max) col(stat) format*”
 - Table format with univariate descriptive statistics that I like best

51

Stata: tabulate

- “tabulate *var*”
 - Provides frequency for each value of a single variable (counts, proportions, cumulative proportion)

52

Stata: Dichotomizing data

- Create a new variable using “generate” and “replace”
- Example: Dichotomize age at 9 and older
 - `g age9over= 0`
 - `replace age9over= 1 if age>=9`

53

Stata: Computing the Mode

- Sample Mode: Discrete data
 - “table *var*”
 - Examine (possibly lengthy) output for highest frequency
- Mode of density for continuous data
 - Graphical display using “*kdensity var*”
 - Better labeling of axes with options “*xlabel*”, “*ylabel*”
 - Numerical output by examining generated variables “*kdensity var, g(x d)*” for the value of *x* that corresponds to maximal value for *d*

54