Biost 517
Applied Biostatistics I

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

Lecture 5:
Descriptive Measures of Spread

October 12, 2009

1

Lecture Outline

- Univariate Measures of Spread
- Univariate Measures of Skewness
- Univariate Measures of Tendency to Extreme Values
- Depictions of Entire Distribution
- Example: Prognostic value of PSA

2

Univariate Measures of Spread

3

Range

- Definition varies:
  - Minimum, maximum values
  - Maximum - minimum is used by some people

4

Range: Types of Variables

- Only makes sense for ordered variables
- Not appropriate for censored time to event
  - Instead use Kaplan-Meier curves to estimate survival or censoring distributions

5

Range: Purpose

- Detecting errors in data collection, entry
  - Values out of range
- Materials and Methods
  - Limits of subjects in sample
- Less useful for quantifying or comparing distributions

6

Range: Scientific Questions

- Scientific questions
  - Not useful unless range of possible values differs across populations
    - But even then, the sampling distribution of the min and max depends quite heavily on the sample size

7

Minimum: Sampling Distribution

- Minimum of n independent and identically distributed random variables

$$\Pr\left(X_{(1)} \geq x\right) = \left(\Pr\left(X \geq x\right)\right)^n$$

  - Tends to estimate the 1/(n+1)-th quantile of the distribution of X
    - 25th %ile when n = 3
    - 1st %ile when n = 99

8

Maximum: Sampling Distribution

- Maximum of n independent and identically distributed random variables

$$\Pr\left(X_{(n)} \le x\right) = \left(\Pr\left(X \le x\right)\right)^n$$

  - Tends to estimate the n/(n+1)-th quantile of the distribution of X
    - 75th %ile when n = 3
    - 99th %ile when n = 99

9

Interquartile Range (IQR)

- Definition varies:
  - 25th, 75th percentiles of sample
  - Difference between quartiles is used by some people

10

IQR: Type of Variables

- Only makes sense for ordered variables
- Not appropriate for censored time to event
  - Estimate quantiles using Kaplan-Meier

11

IQR: Purpose

- Materials and Methods: Characterizing the sample
  - A measure of spread less sensitive to outliers
    - Central 50% of the data
- Assessing validity of assumptions
  - Check for equal spread of distributions
    - BUT: most assumptions are about variances

12

IQR: Scientific Uses

- Quantifying or comparing distributions
  - Sometimes we are scientifically interested in the spread of the distribution
  - The sample quartiles consistently estimate the population quartiles
    - Central tendency not based on sample size

13

Variance

- Definition
  - The average squared distance from the mean

population $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - E(X)\right)^2$$

sample $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

14

Variance: Interpretation

- Population variance has theoretical basis as second central moment of distribution
  - Squared error as a measure of spread
    - Squaring accentuates errors
    - More convenient mathematically
- Relevance to sampling distributions used in statistical inference
  - Variance is a fundamental parameter of the normal distribution
  - Sample variance is an unbiased estimator

15

Variance: Types of Variables

- Variance is a mean
  - Best used with numeric variables having interpretable differences
    - BUT it will be used whenever we are comparing means of distributions
  - Problematic with censored measurements
    - Times are mixture of times to event and times to censoring
    - Indicators of event are measured over varying times

16

Variance: Purpose

- Materials and Methods: Characterizing the spread of the distribution
  - Larger variance means more variable measurements
  - But units are squared units of observations
    - E.g., variance of age is measured in years squared
  - And sensitive to outliers
- Assessing validity of models
  - Many analysis methods rely on assumptions about within group variances

17

Variance: Scientific Questions

- Quantifying or comparing distributions
  - Sometimes we are scientifically interested in the spread of the distribution
  - As a mean, the sampling distribution of the variance in large samples is known
    - But it does take larger sample sizes than is required for inference about the mean

18

Standard Deviation (SD)

- Definition
  - The square root of the variance

population $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - E(X)\right)^2}$

sample $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$

19

SD: Interpretation

- Related to variance; same units as mean
  - Sample SD is not unbiased for population SD
- “Width” of the distribution
  - For any distribution
    - At least 89% of data within 3 SD of mean
  - For the **normal (Gaussian)** distribution
    - About 2/3 of data is within 1 SD of mean
    - About 95% of data is within 2 SD of mean
    - About 99.7% of data is within 3 SD of mean

20

SD vs Variance: Uses

- Variance is just the squared SD
  - Use SD descriptively
    - Units are the same as the measurements
    - Can evaluate equality of variances for assumptions
      - If SDs are equal, then so are variances
  - Use variance for inference
    - Mathematics and distributional theory is better defined for the variance
- SD used for standardization of statistics when making inference about means

21

SD: Standardized Statistics

- Often we measure distance of data from mean in units of SD's
  - (sometimes called "normalized", but does not guarantee normality of data)

$$X_1, \ldots, X_n \sim \left(\text{mean } \mu, \text{variance } \sigma^2\right)$$

$$Z_i = \frac{X_i - \mu}{\sigma}$$

22

Mean Deviation

- Definition
  - The average absolute distance from the mean

$$\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$$

23

Mean Deviation: Uses

- By types of variables
  - Mean deviation is a mean
    - Best for numeric variables having interpretable differences and measured without censoring
- By purpose of descriptive statistics
  - Possible alternative to SD / variance, however
    - Absolute values are harder to work with in calculus
    - Sampling distribution is thus harder to derive
    - Not helpful in statistical inference

24

Univariate Measures of Skewness

Coefficient of Skewness

- Definition
  - The average cubed distance from the mean divided by cube of the standard deviation

sample $$\frac{1}{(n-1)s^3}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3$$

Skewness: Interpretation

- Cubing the distance from the mean
  - Accentuates outliers
  - Does allow positive outliers to cancel out negative outliers
- Symmetric distributions will have a skewness coefficient of 0

Skewness: Types of variables

- Skewness is a mean
  - Best used with numeric variables having interpretable differences
  - Not of interest with censored random variables

Skewness: Purpose

- Materials and Methods: Characterizing the distribution
  - Describing tendency to outliers (in one direction)
- Assessing validity of assumptions
  - Sometimes need symmetric distributions
  - Distributions with outliers generally require larger sample sizes for accurate inference
  - Outliers are sometimes too influential

29

Other Measures of Skewness

- I rarely (never?) compute the coefficient of skewness
  - Usually only interested in qualitatively describing tendency to large (or small) outlying values
- I just use other descriptive statistics to judge possibility of outliers
  - Mean, SD, min, p25, med, p50, max
  - Look at histogram if indicated

30

Symmetric distributions

- Properties of symmetric distributions
  - Mean is equal to the median
  - The median is midway between the minimum and maximum
  - The 25th and 75th percentiles are equidistant from the median

31

Signs of Skewed Distributions

- Descriptive statistics which suggest skewed distributions (especially when due to outliers)
  - The sample median is
    - markedly different from the sample mean
      - (Mean is greatly affected by outliers)
    - not midway between the minimum and maximum
    - not midway between the 25th and 75th percentiles

32

Signs of Skewed Distributions

- An additional criterion is based on the properties of the standard deviation
  - If about 2 standard deviations away from the mean includes impossible values, then it is often the case that large outliers exist
  - E.g., for measurements that must be positive, a standard deviation greater than one-half the mean may suggest the presence of outliers

33

Univariate Measures of Tendency to Extreme Data

34

Coefficient of Kurtosis

- Definition of kurtosis
  - The average fourth power distance from the mean divided by the square of the variance
    - Coefficient of kurtosis subtracts 3

$$\frac{1}{(n-1)s^4}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4 - 3$$

35

Kurtosis: Interpretation

- The fourth central moment will accentuate observations in the tail of the distribution
- In the normal distribution, the fourth central moment is $3\sigma^4$
  - Coefficient of kurtosis is 0

36

Kurtosis: Types of Variables

- Kurtosis is a mean
  - Best used with numeric variables having interpretable differences
  - (As usual,) variables measuring censored times to events can not be described appropriately using the sample coefficient of kurtosis

37

Coefficient of Kurtosis: Uses

- By purpose of descriptive statistics
  - Characterizing the distribution
    - Describing tendency to heavy tails
  - Assessing validity of assumptions
    - The heavier the tails, the larger the sample size needed before the central limit theorem is a reasonable approximation
    - BUT, heavy symmetric tails often shows up in samples as skewness
      - I just tend to look for skewness

38

Characterizing the Entire Distribution

General Comments

39

Statistical Setting

- Statistical analysis of a sample
  - Usually want to make inference to population
- Probability model
  - Describes variation observed in a population
  - We regard that the data were sampled from that probability model

40

Probability Model Classification

- The type of measurement is either
  - Discrete: a countable number of different values are possible for the measurement, or
  - Continuous: an uncountable number of different values are possible for the measurement

41

Samples vs Populations

- All (finite) samples are discrete
  - Descriptive measures appropriate for discrete measurements always make sense
- Often, however, we will be trying to estimate quantities that are only appropriate for continuous measurements (e.g., densities)

42

Describing Discrete Distns

- Typically give probability of each outcome
  - Prob mass function (pmf):Pr (X = x) for each x
    - Explicit numbers
      - (only possibility for unordered categorical data)
    - Formulas
      - (especially for measurements representing counts)
- Ordered variables: we can also consider the cumulative distribution function (cdf):
    - Pr (X ≤ x) for every x

43

Describing Continuous Distns

- Typically give a function that can be used to define the probability that the random variable is in some interval
  - Density (pdf)
    - Similar to Pr (X = x)
      - (but must be integrated over an interval)
  - Cumulative distribution function (cdf)
    - Pr (X ≤ x) for every x
  - Survivor function
    - Pr (X > x) for every x

44

Why Describe Entire Distribution

- Assessing validity of data
  - Viewing outliers
- Quantifying distributions within groups
  - Simultaneously consider location, spread
- Assessing validity of assumptions for modeling
  - Outliers, shape of distribution
- Hypothesis generation
  - Multimodality, etc.

45

Tabling a Distribution

46

Frequency Table

- Frequency or proportion for each possible value of a discrete variable
  - Not defined for a continuous measurement in a population, but always defined for a sample
  - With unordered categorical data this is the most logical summary
  - With a sample from a continuous variable, this makes most sense when data is grouped into intervals
    - E.g., age divided into decades

47

Frequency Table: Issues

- Missing data
  - Must consider whether missing data should be counted as part of the denominator or not
- Order of listing categories
  - Unordered data: alphabetical versus most frequent, etc.
  - Ordered data: usually use ordering

48

Stata Ex: Bone Scan Score

- `tabulate bss`

bss	Freq.	Percent	Cum.
1	5	10.42	10.42
2	13	27.08	37.50
3	30	62.50	100.00
Total	48	100.00	

49

Stata Ex: Bone Scan Score

- `tabulate bss, missing`

bss	Freq.	Percent	Cum.
1	5	10.00	10.00
2	13	26.00	36.00
3	30	60.00	96.00
.	2	4.00	100.00
Total	50	100.00	

50

Frequency Table: Issues

- Categorization of continuous data
  - Number of groups
    - Tradeoffs between finer resolution and lots of groups with zero counts
    - Based on average counts per group?
      - » log base 2 of N?
  - Width of intervals
  - Cutpoints
    - Based on scientific interest
    - Based on number of observations
      - » E.g., intervals based on quintiles

51

Categorizing Continuous Data

- My recommendations
  - Groups based on scientific considerations
    - E.g., Age by decades rather than quintiles
  - Number of groups
    - Reasonable sample sizes for reliable estimates
    - Space constraints in tables

52

Stata: Categorizing Data

- Categorizing continuous random variables
  - "recode var rulelist"
    - Replaces values of the variable
    - Rules of the form (for variable x)
      - min/3=1 (changes $x \leq 3$ to 1)
      - 4=2 (changes 4 to 2)
      - 5/9=3 (changes $5 \leq x \leq 9$ to 3)
      - 10/max=4 (changes $x \geq 10$ to 4)
    - Each rule contains both endpoints of range, last rule listed is used to resolve conflicts
    - Labels can be assigned using "label"

53

Stata Ex: Age

- `tabulate age`

age	Freq.	Percent	Cum.
58	2	4.00	4.00
61	5	10.00	14.00
62	1	2.00	16.00
63	5	10.00	26.00
64	5	10.00	36.00
65	2	4.00	40.00
66	6	12.00	52.00
68	7	14.00	66.00
69	3	6.00	72.00
70	2	4.00	76.00
71	4	8.00	84.00
73	1	2.00	86.00
74	1	2.00	88.00
75	2	4.00	92.00
78	1	2.00	94.00
79	1	2.00	96.00
81	1	2.00	98.00
86	1	2.00	100.00
Total	50	100.00	

54

Stata Ex: Categorizing Age

- `g agectg = age`
- `recode agectg min/60=1 60/70=2 70/80=3 80/max=4`
- `tabulate agectg`

```
 agectg |      Freq.     Percent        Cum.
--------+-----------------------------------
      1 |          2        4.00        4.00
      2 |         34       68.00       72.00
      3 |         12       24.00       96.00
      4 |          2        4.00      100.00
--------+-----------------------------------
  Total |         50      100.00
```

55

Graphing a Distribution

56

Stem-leaf Plot

- Stem-leaf plot : Intermediate to a tabular listing of the data and a histogram
  - Only appropriate for ordered quantitative data
  - Construction of the stem-leaf plot
    - Each measurement is divided into
      - a “stem” by truncating the observation in some position, e.g., tens digit
      - a “leaf” the remaining value
      - (thus the measurement is equal to the “stem”+”leaf”)

57

Stem-leaf Plot

- Construction of the stem-leaf plot (cont.)
  - The rows are the ordered “stem”s
  - In each row, the first digit of all the “leaf”s for that stem are listed (ordered or unordered)
- Advantages
  - Graphical appearance of a histogram
  - Retains more information about the individual measurements
  - Can be made “on the fly”
- Stata Command: “stem”

58

Stata Ex: Stem-Leaf Plot of Age

```
. stem age
     5. | 88
     6* | 11111
     6t | 233333
     6f | 4444455
     6s | 666666
     6. | 8888888999
     7* | 001111
     7t | 3
     7f | 455
     7s |
     7. | 89
     8* | 1
     8t |
     8f |
     8s | 6
```

59

S-Plus: Stem-Leaf Plot of Age

Decimal point is 1 place to the right of the colon

5 : 88
6 : 1111123333344444
6 : 5566666688888888999
7 : 00111134
7 : 5589
8 : 1

High: 86

60

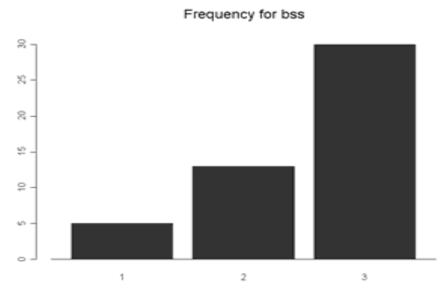
Bar Plots

- Frequencies for categorical data
  - A separate bar for each category
  - (Not exactly the same as a histogram, which is used to estimate a density)

61

Ex: Bone Scan Score (S-Plus)

- Bar plot of bone scan score in PSA data



62

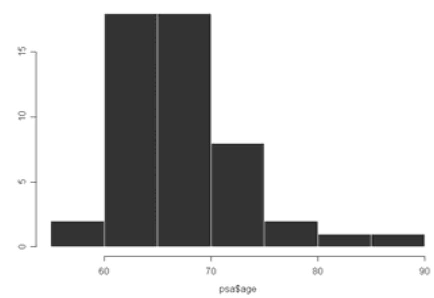
Histograms

- A plot of the frequency of (categorized) continuous measurements
  - Tabulate counts of each (grouped) measurement
  - Plot bars for each group
    - Width is width of interval
    - Height such that area is proportional to the count
      - (wider intervals should decrease the height accordingly)

63

Ex: Histogram of Age (S-Plus)

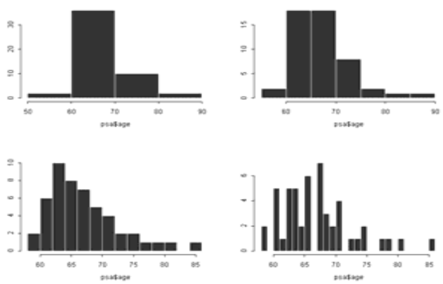
- Histogram of age in PSA dataset



64

Ex: Histograms of Age (S-Plus)

- Histograms with varying number of groups can look quite different for the same data



65

Histogram: Comments

- Histograms are attempting to estimate a density
- The appearance of a histogram can be quite variable depending upon the selection of the groups
  - Number of groups
  - Cutpoints of groups
- Density estimation is probably better

66

Density Estimation

- Density estimates are essentially smoothed histograms
  - Only make sense for continuous measurements
- Smoothers can be used to provide better estimates of the density
  - In a kernel smoother, each point is “distributed” over a range of measurements
  - The “kernel” describes how many adjacent measurements are used to estimate the density and how the points are weighted

67

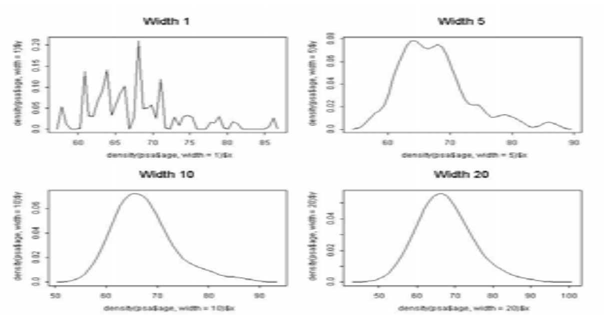
Stata: Density Estimates

- “kdensity var, (options)”
  - Options include
    - Shape of kernel (biweight, cosine, etc.)
    - Width of window
      - » how distant a point can influence density estimate
    - Number of points estimated
- (I tend to use default values for the options)

68

Ex: Density Estimation

- Age from PSA data set (Gaussian window, varying widths)



69

Cumulative Distribution, Survivor Graphs

- For ordered variables
  - Used to estimate the corresponding quantity for a population
    - These functions can sometimes be estimated (and graphed) for censored data (unlike histograms, densities, etc.)
  - We will discuss these graphs further next lecture

70

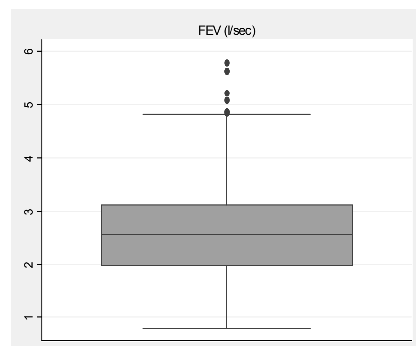
Box Plots

- Display several summary measures simultaneously
  - A box is drawn from the lower quartile to the upper quartile, with a dividing line drawn at the median
  - Whiskers are either
    - min and max (in the absence of “outliers”), or
    - limits of “nonoutlying” data (as defined by an arbitrary criterion)
  - “Outliers” are plotted separately

71

Box Plots: FEV

- `graph box fev, t1(“FEV(l/sec)”)`



72

Example: Use of Univariate Descriptive Statistics

73

Standard Univariate Description

- Often easier to just ask for standard descriptive statistics on all variables
  - Sample size
  - Number of missing
  - Mean
  - Standard Deviation
  - Minimum
  - 25th percentile
  - Median (50th percentile)
  - 75th percentile
  - Maximum

74

Standard Univariate Description

- We must then consider how to use the relevant statistics (and how to ignore the irrelevant ones)
  - Scientific relevance
  - Descriptive measures
    - Measures of location
    - Measures of spread
    - Detecting outliers

75

Example:

- Prognostic value of PSA in hormonally treated prostate cancer
  - Usefulness of nadir PSA in predicting time in remission
  - Type of question?
    - Comparing distributions?
    - Predicting time in remission?

76

Example: PSA Data Variables

- Prognostic value of PSA in hormonally treated prostate cancer
  - ptid Patient ID
  - nadir Lowest PSA following treatment
  - pretx Pre-treatment PSA
  - ps Performance status (0 – 100)
  - bss Bone scan score (1, 2, or 3)
  - grade Tumor grade (1, 2, or 3)
  - age Age (years)
  - obstime Time until relapse or end of study (months)
  - inrem In remission at obstime

77

Ex: PSA Descriptive Statistics

	n	ms	mean	stdev	min	25%le	mdn	75%le	max
ptid	50	0	25.5	14.6	1.0	13.2	25.5	37.8	50
nadir	50	0	16.4	39.2	0.1	0.2	1.0	9.5	183
pretx	50	7	670.8	1287.6	4.8	52.0	127.0	408.0	4797
ps	50	2	80.8	11.1	50.0	80.0	80.0	90.0	100
bss	50	2	2.5	0.7	1.0	2.0	3.0	3.0	3
grade	50	9	2.2	0.8	1.0	2.0	2.0	3.0	3
age	50	0	67.4	5.8	58.0	63.2	66.0	70.0	86
obstime	50	0	28.5	18.4	1.0	12.5	28.0	42.0	75
inrem	50	0	0.3	0.4	0.0	0.0	0.0	1.0	1

78

Example: PSA Data Variables

- Types of data
  - ptid Unordered categorical (coded as numbers)
  - nadir Continuous (ratio)
  - pretx Continuous (ratio)
  - ps Continuous (ratio) (measured discretely)
  - bss Ordered categorical
  - grade Ordered categorical
  - age Continuous (ratio)
  - obstime Censored continuous
  - inrem Binary indicator of censoring for obstime

79

Example: PSA Data Variables

- Relevant univariate statistics
  - ptid (Mode: are observations independent?)
  - nadir Mean, SD, Min, Max, Quantiles
  - pretx Mean, SD, Min, Max, Quantiles
  - ps Mean, SD, Min, Max, Quantiles
  - bss Min, Max, Quantiles (Frequencies)
  - grade Min, Max, Quantiles (Frequencies)
  - age Mean, SD, Min, Max, Quantiles
  - obstime (Kaplan-Meier estimates needed)
  - inrem (Kaplan-Meier estimates needed, though mean of inrem does tell us proportion of uncensored observations but not time)

80

Example: Relevant Univariate Statistics

	n	ms	mean	stdev	min	25%le	mdn	75%le	max
ptid	50	0							
nadir	50	0	16.4	39.2	0.1	0.2	1.0	9.5	183
pretx	50	7	670.8	1287.6	4.8	52.0	127.0	408.0	4797
ps	50	2	80.8	11.1	50.0	80.0	80.0	90.0	100
bss	50	2			1.0	2.0	3.0	3.0	3
grade	50	9			1.0	2.0	2.0	3.0	3
age	50	0	67.4	5.8	58.0	63.2	66.0	70.0	86
obstime	50	0							
inrem	50	0							

81

Example: PSA Data Skewness

- Detecting skewness
  - Both nadir and pretx appear markedly skewed
    - Mean, median markedly different
    - Median not midpoint of range
    - Median not midpoint of interquartile range
    - SD greater than one-half of mean for these positive measurements
  - No evidence of skewness (that I care about) for age or pss

82

Ex: Bone Scan Score

- `tabulate bss`

bss	Freq.	Percent	Cum.
1	5	10.42	10.42
2	13	27.08	37.50
3	30	62.50	100.00
Total	48	100.00	

83

Ex: Bone Scan Score

- `tabulate bss, missing`

bss	Freq.	Percent	Cum.
1	5	10.00	10.00
2	13	26.00	36.00
3	30	60.00	96.00
.	2	4.00	100.00
Total	50	100.00	

84

Ex: Categorizing Age

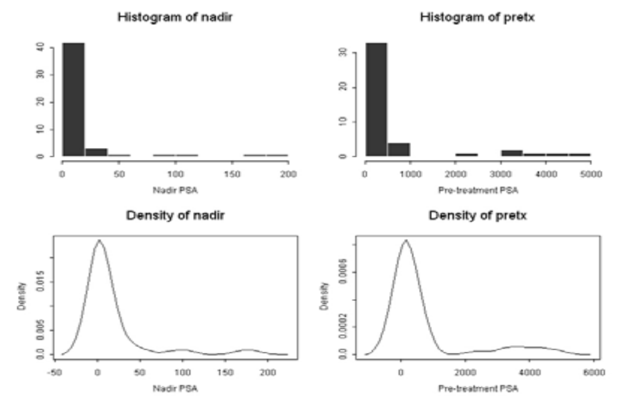
- g agectg = age
- recode agectg min/60=1 60/70=2 70/80=3 80/max=4
- tabulate agectg

```
 agectg |      Freq.     Percent        Cum.
--------+-----------------------------------
      1 |          2        4.00        4.00
      2 |         34       68.00       72.00
      3 |         12       24.00       96.00
      4 |          2        4.00      100.00
--------+-----------------------------------

  Total |         50      100.00
```

85

Example: PSA Data Variables



86

Example: PSA Data Variables