Biost 517
Applied Biostatistics I

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

Lecture 7:
Bivariate Descriptive Statistics

October 21, 2009

1

Lecture Outline

- Review of univariate descriptive statistics
- Purpose of bivariate descriptive statistics
- Stratified univariate descriptives
- Graphical
  - Stratified histograms, densities, boxplots
  - Scatterplots, least squares lines, smooths
- Correlation

2

Review of Univariate Descriptive Statistics

3

Univariate Descriptive Statistics

			Unordered	Ordered		
		Binary	Nominal	Categ	Quant	Cens
Entire Distnution	**Frequency**	OK	OK	OK	OK	
	Cum Freq	**boring**		OK	OK	KM
	Mode	**boring**	Sample	Sample	Density	
	Min / Max	**boring**		**boring**	OK	
Dicho-tomize	**Proportion (or Odds)**	OK	OK	OK	OK	KM
Quant-iles	**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)

Purpose of Bivariate Descriptive Statistics

Purpose of Bivariate Description

- Characterize the relationship between two variables
  - Detecting errors in data collection or data entry
  - Characterizing materials and methods
  - Assessing validity of assumptions
  - Basis for some estimates of association (inference)
  - Hypothesis generation (exploration/inference)

Detecting Errors

- Sometimes data measurements are not unusual univariately, but are unusual in combination
  - E.g., 6 foot tall 3 year olds
  - E.g., pregnant males
- Patterns may exist in missing data
  - Minorities may be more likely to be missing data on some medical procedures

Materials and Methods

- Describe patterns of sampling
  - E.g., minorities may tend to be younger
  - E.g., older subjects may tend to be women
  - E.g., smokers may tend to drink alcohol more

Validity of Assumptions

- Scientific
  - Confounding (ultimately involves 3 variables)
  - Effect modification (involves 3 variables)
- Statistical
  - E.g., assumptions about within group variance
  - E.g., assumptions about linearity of trends
  - E.g., influence of "outliers"

9

Evidence of Associations

- Two variables are said to be associated if the distribution of one variable differs across groups defined by the other variable
  - E.g., if interested in determining whether sex and blood pressure are associated, see if
    - distribution of blood pressure differs between men and women, OR
    - proportion of men varies across groups defined by blood pressure measurements

10

Quantify Associations

- Describe "dose-response"
  - How the effect differs across groups having ever larger differences in the grouping variable
    - E.g., linear response
    - E.g., S-shaped curves
    - E.g., U-shaped trends

11

Hypothesis Generation

- Examine sample to detect associations not previously considered
  - Any such associations suggested by the data should be confirmed in an independent sample

12

Stratified Univariate Descriptive Statistics

Stratified Univariate Descriptives

- Strata defined by one variable
  - Continuous variables must be divided into categories
  - Methods of dividing into categories
    - Scientific basis: Intervals with scientific meaning
    - Statistical basis: Intervals with equal sample sizes
      - Generally such intervals are not evenly spaced, thus detecting linear trends are difficult

Examining Stratified Descriptives

- Errors in data:
  - Unusual range by strata
- Materials and methods:
  - Describe central tendency, range by strata
- Validity of assumptions:
  - E.g., missing data by strata
  - E.g., equal variances across strata
  - E.g., linear trends in central tendency
- Evidence of associations:
  - E.g., difference in means, medians across strata

Choice of Stratified Descriptives

- By purpose of descriptives
  - Ability to assess errors
  - Ability to describe sample
  - Ability to examine validity of assumptions
  - Scientific relevance as estimates of association
- By type of data
  - Unordered versus ordered
  - Continuous versus discrete
  - Uncensored versus censored

FEV Ex: Categorizing Age

- Stata: Create variable containing age categories in two year intervals
  - `g agectg = age`
  - `recode agectg 3/4=3 5/6=5 7/8=7 9/10=9 11/12=11 13/14=13 15/16=15 17/18=17 19/20=19`
  - Alternative approach using arithmetic
    - `g agectg = int ( (age-1) / 2) * 2 + 1`

17

FEV Ex: Stata Commands

- Stata: Create table of stratified statistics
  - `by agectg: tabstat height, stat(n mean sd min p25 p50 p75 max) col(stat) format`

18

FEV Ex: Height by Age Groups

agectg	N	mean	sd	min	p25	p50	p75	max
3	11.0	48.8	1.8	46.0	48.0	48.0	50.0	52.0
5	65.0	52.5	2.4	46.5	51.0	52.5	54.0	58.0
7	139.0	57.1	3.3	47.0	54.5	57.0	59.5	67.5
9	175.0	61.5	3.3	52.5	59.0	61.0	64.0	70.0
11	147.0	64.7	3.3	57.0	62.0	64.5	67.0	72.0
13	68.0	66.7	3.5	61.0	63.8	67.0	69.0	74.0
15	32.0	66.8	3.4	60.0	64.0	66.8	69.3	73.5
17	14.0	67.7	3.6	60.0	66.0	68.5	70.0	73.0
19	3.0	67.8	3.6	65.5	65.5	66.0	72.0	72.0
Total	654.0	61.1	5.7	46.0	57.0	61.5	65.5	74.0

19

FEV Ex: Findings

- Mean, median height within age strata increases by about 4 inches every two years up until about age 12 then levels off
- Standard deviation of height within age strata much less than standard deviation of entire sample
  - An indication that age predicts height
- Standard deviation of height within age strata increases as the mean increases

20

Mean-Variance Relationships

- We often see the variance differ systematically according to group means
  - Differential diagnosis
    - Precision of measurement as a percentage
    - Variability in rates, but measurement of total
      - E.g., different growth per year, height at age 10
    - Confounding
      - E.g., more older kids smoke, which stunts growth?
    - Effect modification
      - E.g, young boys and girls tend to be same height, older boys taller than older girls

21

FEV Ex: Age Quantiles

- Stata: Find age quantiles
  - `centile age,c(12 25 37 50 62 75 87)`

```
    Variable | Obs  Pctile  Centile    [95% Conf. Interval]
-------------+------------------------------------------
         age | 654     12       7.0         6.0         7.0
             |         25       8.0         8.0         8.0
             |         37       9.0         9.0         9.0
             |         50      10.0         9.0        10.0
             |         62      11.0        10.0        11.0
             |         75      12.0        11.0        12.0
             |         87      13.0        13.0        14.0
```

22

FEV Ex: Categories by Quantiles

- Stata: Create variable of age categories
  - `g agectg=age`
  - `recode agectg min/7=1 8=2 9=3 10=4 11=5 12=6 13=7 14/max=8`
  - `tabstat fev, stat(n mean sd min p25 p50 p75 max) col(stat) format by(agectg)`

23

FEV Ex: Height by Age Octiles

agectg	N	mean	sd	min	p25	p50	p75	max
1	130.0	53.4	3.1	46.0	51.0	53.0	55.5	62.5
2	85.0	58.3	3.2	52.0	56.5	58.5	60.0	67.5
3	94.0	60.6	2.9	53.0	58.5	60.3	62.5	69.0
4	81.0	62.5	3.4	52.5	60.0	62.0	65.0	70.0
5	90.0	64.5	3.2	58.0	62.0	64.5	67.0	72.0
6	57.0	65.2	3.5	57.0	63.0	64.5	68.0	72.0
7	43.0	66.2	3.6	61.0	63.0	66.5	68.5	74.0
8	74.0	67.2	3.4	60.0	64.5	67.0	70.0	73.5
Total	654.0	61.1	5.7	46.0	57.0	61.5	65.5	74.0

24

FEV Ex: Findings with Quantiles

- We were less able to pick out the regions of linearity
  - The lowest octile covered several years, the next few octiles were each 1 year in width
- We were not able to pick out the mean variance relationship
  - The first octile was not as homogeneous with respect to age, and height varied with age within that stratum

25

Crosstabulation of Categories

- `tabulate smoker female, cell column row`

```
       |        female
smoker |         0          1 |     Total
-------+----------------------+----------
       |       310        279 |       589
       |     52.63      47.37 |    100.00
       |     92.26      87.74 |     90.06
     0 |     47.40      42.66 |     90.06
-------+----------------------+----------
       |        26         39 |        65
       |     40.00      60.00 |    100.00
       |      7.74      12.26 |      9.94
     1 |      3.98       5.96 |      9.94
-------+----------------------+----------
 Total |       336        318 |       654
       |     51.38      48.62 |    100.00
       |    100.00     100.00 |    100.00
       |     51.38      48.62 |    100.00
```

26

Graphical Bivariate Descriptive Statistics

27

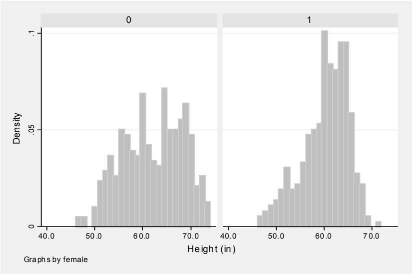
Stratified Univariate Graphs

- Divide sample into strata based on one variable
- Display for each stratum
  - histograms
  - densities
  - boxplots
- For greatest comparability, the same axes should be used in all plots

28

Ex: Stratified Histograms

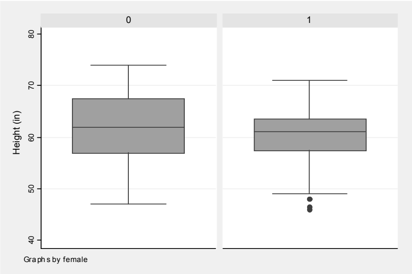
- FEV data: Histograms of height by sex
  - `hist height, by(female) xtitle("Height (in)")`



29

Ex: Stratified Boxplots

- FEV data: Boxplots of height by sex
  - `graph box height, by(female) ytitle("Height (in)")`



30

Ex: Stratified Univariate Stats

- FEV data: Univariate description of height by sex
  - `tabstat height, by(female) stat(n mean sd min p25 p50 p75 max) col(stat) format`

female	N	mean	sd	min	p25	p50	p75	max
0	336.0	62.0	6.3	47.0	57.0	62.0	67.5	74.0
1	318.0	60.2	4.8	46.0	57.5	61.0	63.5	71.0
Total	654.0	61.1	5.7	46.0	57.0	61.5	65.5	74.0

31

Scatterplots

- A graph of Y versus X
  - Most useful for two continuous variables
- Look for
  - Outliers
  - Trends in location across groups
    - First order trends (linear)
    - Second order trends (curves, U-shape, S-shape)
  - Trends in within group spread of data
    - (Looking at range)

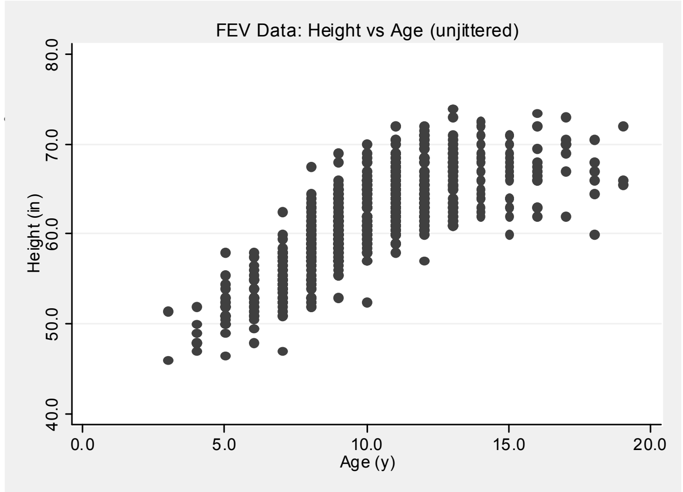
32

Stata: Scatterplot

- Stata commands
  - "scatter *yvar xvar, [options]*"
- Example: Height vs Age in FEV Data

```
scatter height age, xtitle("Age (y)")
t1("FEV Data: Height vs Age (unjittered)")
ytitle("Height (in)")
```

33



34

Ex: Interpretation

- No outliers
- Tends to increased height for older ages
  - First order trend is upward
- Hint of curvilinear relationship
  - Height levels off at highest ages
- Suggestion of increasing spread with increased height
  - Must be careful when judging variability from range
  - Need to compare range of equal numbers of data in area with equal slopes

35

Jittered Scatterplots

- If variables are discretely measured, jittering can be helpful
  - "jittering": adding a little noise to the data to break ties
  - I tend to try to jitter to allow visualization of all points, but still try to keep discrete levels separate use a spread of about 40% the separation between categories

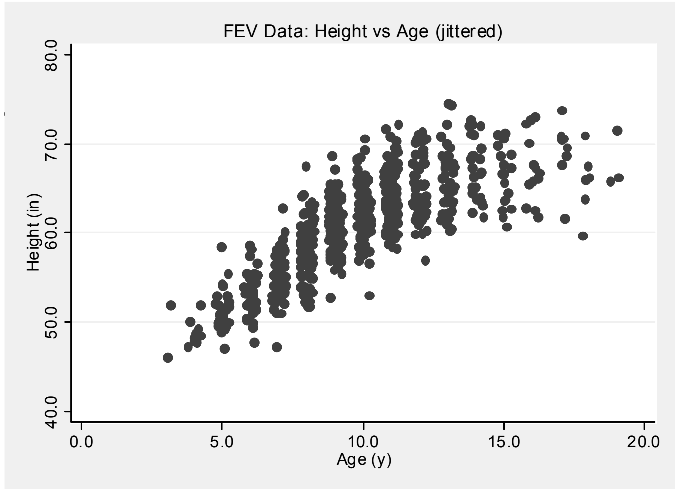
36

Stata: Jittering

- Stata commands
  - "scatter *yvar xvar*, jitter*(n)*"
- Example: Height vs Age in FEV Data

```
scatter height age, xtitle("Age (y)")
t1("FEV Data: Height vs Age (jittered)")
ytitle("Height (in)") jitter(3)
```

37



38

Variance Within Groups

- On a scatterplot, our eye sees range of data within groups
- We usually want to judge variance
  - Especially how variance might differ with X
- Converting range to variance
  - Consider spread in two regions far apart
    - Need sample sizes approximately equal
    - Need slopes approximately equal

39

Superimposed Curves

- It is often helpful to place curves over a scatterplot to help see trends in the data
  - Theoretical relationship
    - If theory prescribes a supposed relationship
  - Least squares line of within group means
    - "Best fitting" line to means, but has to be a line
  - Smooths of measure of location within groups
    - Curve representing approximation to the data
      - E.g., lowess

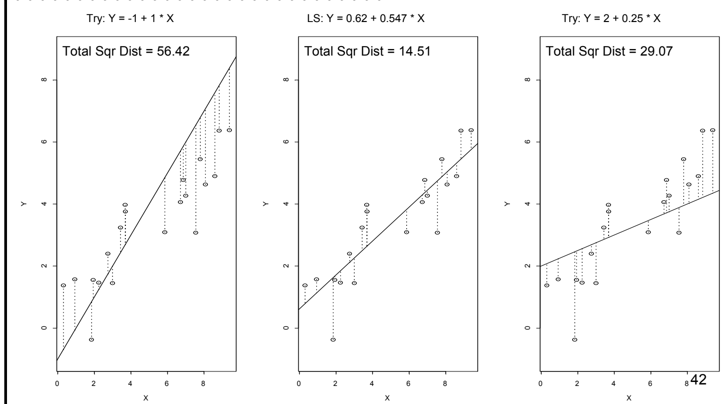
40

Least Squares Line

- Find the straight line that minimizes total squared vertical distance from data to line
  - Conceptually: Trial and error search
    - Guess a formula for a line
    - Compute total squared distance from data to line
    - Iterate until smallest number found
  - Calculus:
    - Find a formula based on derivatives
  - Real life:
    - Computers find such estimates easily

41

Conceptual Example



42

Stata: Superimposed Lines

- Basic Stata bivariate graph command
  - "`twoway …`"
  - Special cases
    - "`twoway scatter …`" (scatterplot of points)
    - "`twoway line …`" (connect with lines)
    - "`twoway lfit …`" (least squares fit)
    - "`twoway lowess …`" (lowess curve)
- Superimposed graphs
  - `twoway (`*`graphtyp …`*`) (`*`graphtyp …`*`) …`

43

Ex: Height vs Age with LS Fit

- Stata commands

```
twoway (scatter height age, jitter(3))
(lfit height age), xtitle("Age (y)")
ytitle("Height (in)")
t1("Least Squares Fit of Height on Age")
```

44

Ex: Height vs Age with LS Fit

45

Interpretation

- Clearly increasing trend in data
- Our eye tends to like to detect lines, so it takes careful inspection to decide a line is not the best fit
  - Note that at lowest ages and highest ages most data tend to be on one side of line rather than symmetric about line
  - Possible curvilinear association

46

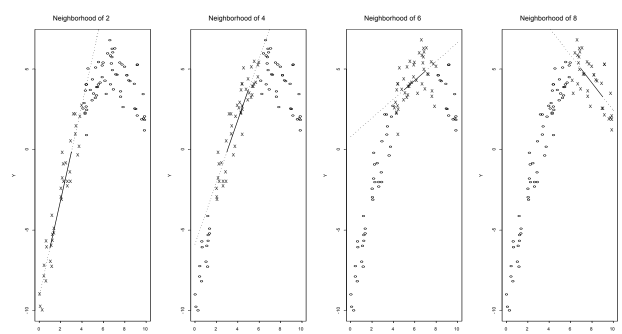
Lowess Smooths

- Locally Weighted Scatterplot Smoother
  - A smoother to find a smooth curve approximating relationship in the data
  - For every value of X, fits straight lines in a neighborhood of that value
    - "Bandwidth" is width of window defining neighborhood
    - Weights closer data more heavily
  - Combines the estimates from different regions to form a smooth curve

47

Lowess: Conceptual Algorithm

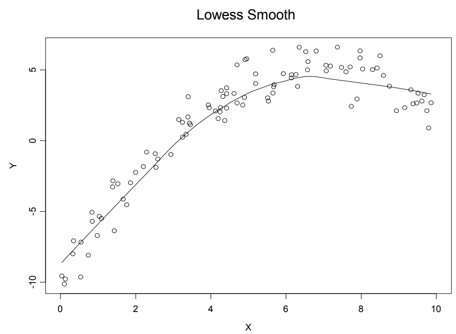
- Least squares lines fit in neighborhoods



48

Lowess Smooth

- Combines locally fit least squares lines

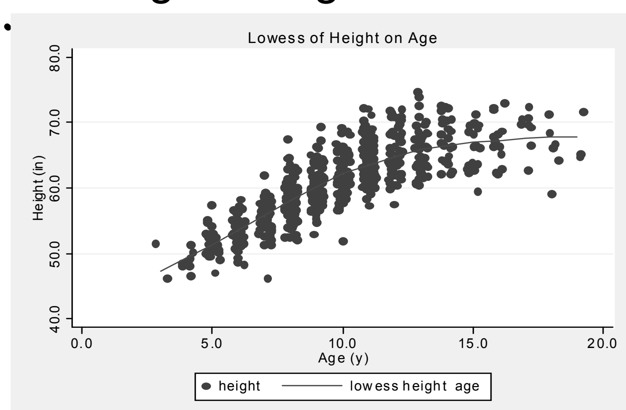


Ex: Height vs Age with Lowess

- Stata commands

```
twoway (scatter height age, jitter(3))
(lowess height age),
xtitle("Age (y)") ytitle("Height (in)")
t1("Lowess of Height on Age")
```


Ex: Height vs Age with Lowess

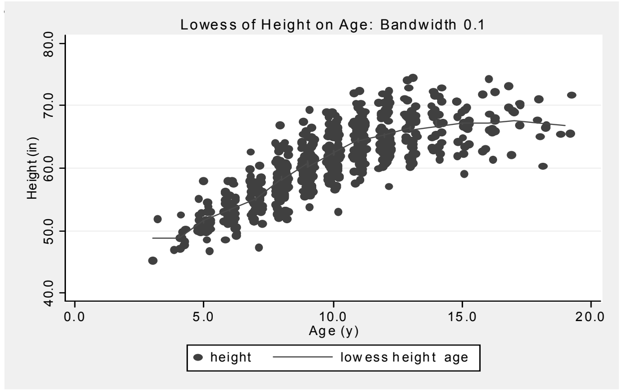


Changing the Bandwidth

- Default bandwidth is 0.8 (80% of data)
  - I typically use the default of whatever program I am using
- Stata commands for less smoothing

```
twoway (scatter height age, jitter(3))
(lowess height age, bwidth(.1)),
xtitle("Age (y)") ytitle("Height (in)")
t1("Lowess of Height on Age: Bandwidth 0.1")
```

Changing the Bandwidth



53

Ex: Showing Both LS, Lowess

- Stata commands

```
twoway (scatter height age, jitter(3))
(lowess height age, col("red")
(lfit height age, col("blue"),
xtitle("Age (y)") ytitle("Height (in)")
t1("Lowess, LS of Height on Age")
```

54

Ex: Showing Both LS, Lowess

55

Interpretation

- Lowess smooth shows that height tends to increase pretty linearly with age up until about age 11 or 12
- Height levels off in late teens with little change in mean height

56

Other Smoothers

- Many different methods of smoothing data have been proposed
  - Lowess is often criticized due to the way it can accentuate data near the end of its range
    - One should not make too much of the way the estimate curve wiggles at the extremes of the data
  - For my purposes, almost any smoother will do
    - I just want to have something that is not forced to be a line, and something that I did not draw
      - I can be just as biased as anyone

57

Correlation

58

Correlation Coefficient

- A measure of the tendency of the largest measurements for one variable to be associated with the largest measurements of the other variable
  - Dimensionless
  - The sample correlation r estimates the population correlation ρ (rho)

59

Pearson's Correlation Coefficient

•Definition of correlation between *X* and *Y*:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}\sqrt{\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2}}$$

60

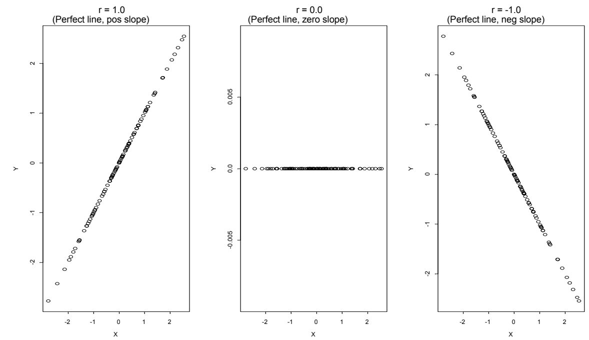
Possible Values of r

- Range of r : $-1 \leq r \leq 1$
  - r = 1 : perfect positive correlation
    - a graph of X vs Y will be a straight line with positive slope
  - r = -1 : perfect negative correlation
    - a graph of X vs Y will be a straight line with negative slope
  - r = 0 : no correlation

61

Straight Line Relationships

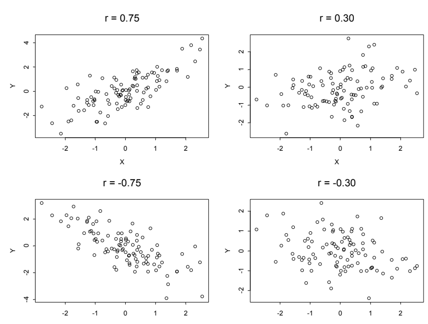
- Pearson's correlation coefficient with linear data



62

Linear Trends in Data

- Pearson's correlation coefficient with variable data



63

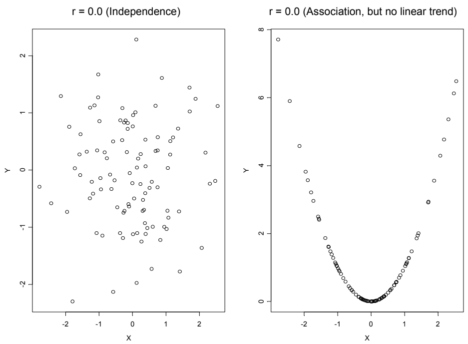
Correlation and Independence

- Independent variables will have $\rho = 0$
  - (and r tending to be close to 0)
- However, uncorrelated variables are not necessarily independent
  - Correlation measures linear trend in the mean of one variable in groups defined by the other
  - It is possible that a nonlinear association exists between two variables, and that the first order trend is a zero slope

64

Uncorrelated Variables

- No linear trend between the variables



65

Stata Commands

- "`correlate` *varlist*"
  - Correlation of all pairs of variables
  - Missing data deleted on a casewise basis
- "`pwcorr` *varlist*"
  - Correlation of all pairs of variables
  - Missing data deleted on a pairwise basis

66

Ex: Correlation in FEV Data

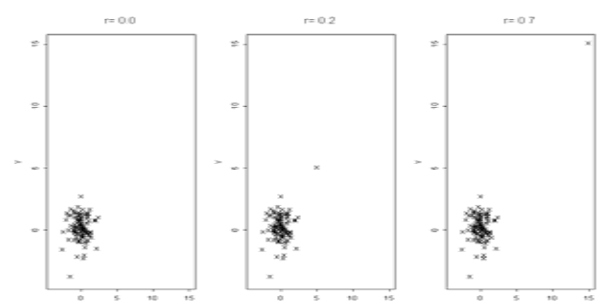
```
. corr subjid age fev height sex smoke
(obs=654)
       | subjid     age     fev  height     sex   smoke
-------+------------------------------------------------
subjid | 1.0000
   age |-0.0112  1.0000
   fev |-0.0147  0.7565  1.0000
height |-0.0317  0.7919  0.8681  1.0000
   sex | 0.0407 -0.0291 -0.2084 -0.1590  1.0000
 smoke |-0.0601 -0.4043 -0.2454 -0.2804 -0.0756  1.0000
```

- Some of these correlations don't make much sense
  - subjid is a nominal variable
  - sex, smoke are binary variables

67

Effect of Outliers on r

- Pearson's correlation coefficient can be greatly affected by outliers



68

Spearman's Rank Correlation

- To decrease the influence of outliers, Spearman's rank correlation coefficient computes the correlation of the ranks of the data
  - In the previous example, the rank correlation is always the same: approximately 0.07

69

Stata: Spearman's Correlation

- "**`spearman var1 var2`**"
  - Correlation of one pair of variables
  - Cases with missing data for either variable are deleted, and then ranks are computed

70

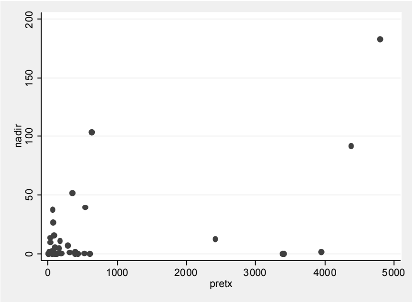
Ex: Correlation in PSA Data

```
corr nadir pretx
(obs=43)
        |    nadir    pretx
--------+------------------
   nadir|   1.0000

   pretx|   0.5371   1.0000


spearman nadir pretx
Number of obs =      43
Spearman's rho =   0.1489
```

71

Ex: Nadir vs Pretreatment PSA

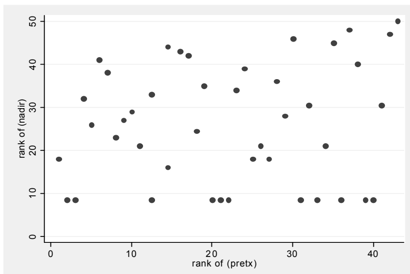
- Scatterplot of nadir versus pretx
  - scatter nadir pretx



72

Ex: Nadir vs Pretx Ranks

- egen rnknadir = rank(nadir)
- egen rnkpretx = rank(pretx)
- scatter rnknadir rnkpretx



73

Ex: Spearman's Corr vs r

- Possible explanation for lower rank correlation with Spearman's
  - Perhaps outliers in distribution of nadir and/or pretx unduly inflate r
  - Perhaps transforming to ranks masks true linear association in skewed variables

74

Uses of Correlation

- By type of variable
  - Correlation is a mean, thus only makes sense when a mean does
    - Limited interpretability with categorical data
    - Of no scientific relevance with censored data
- By scientific question
  - Greatest relevance when looking for associations between variables
    - But not particularly generalizable across studies

75

Correlation and Regression

76

More Interpretable Formula for r

$$r \approx \beta \sqrt{\frac{Var(X)}{\beta^2 Var(X) + Var(Y \mid X = x)}}$$

β = (LS) slope between Y and X

$Var(X)$ = variance of X in sample

$Var(Y \mid X = x)$ = variance of Y in groups that have same value of X

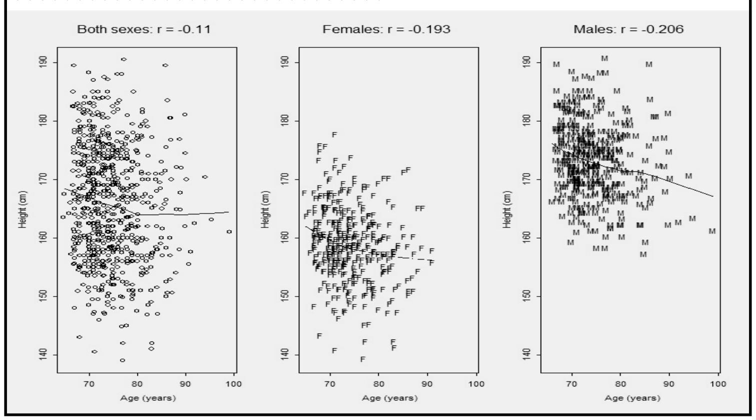
77

Properties of Correlation

- Correlation tends to increase in absolute value as
  - The absolute value of the slope of the line increases
  - The variance of data decreases within groups that share a common value of X
  - The variance of X increases
  - (Sample size is unimportant in tendencies toward lower or higher correlation)

78

Ex: Height vs Age (by Sex)

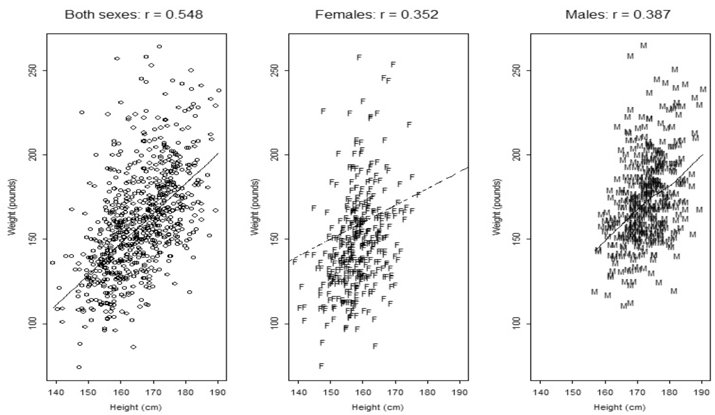


Ex: Height vs Age (by Sex)

- Correlation between Height and Age
  - Males: r = -0.206; Females: r = -0.193
  - Combined: r = -0.110
- Less extreme r in combined sexes
  - Approximately same slope in each sex and overall
  - Approximately same variance of age in each sex and overall
  - Combined group has higher within group variance of height by age (due to sex effect)

80

Ex: Weight vs Height (by Sex)



Ex: Weight vs Height (by Sex)

- Correlation between Height and Weight
  - Males: r = .387; Females: r = 0.352
  - Combined: r = 0.548
- More extreme r in combined sexes
  - Approximately same slope in each sex and overall
  - Approximately same within group variance (by height) for each sex and overall
  - Combined group has higher variance of height

82

Scientific Relevance of r

- Correlation as a scientific measure
  - It should be noted that
    - the slope between X and Y is of scientific interest
    - the variance of Y|X=x is partly of scientific interest, but it can be affected by restricting sampling to certain values of another variable
      - E.g., var (Height | Age) is less in males than when both sexes are included
    - the variance of X is often set by study design
      - This is often not of scientific interest

83