3. Descriptive Statistics

- Describing data with *tables* and *graphs* (quantitative or categorical variables)
- Numerical descriptions of *center*, *variability*, *position* (quantitative variables)
- *Bivariate* descriptions (In practice, most studies have *several* variables)

1. Tables and Graphs

Frequency distribution: Lists possible values of variable and number of times each occurs

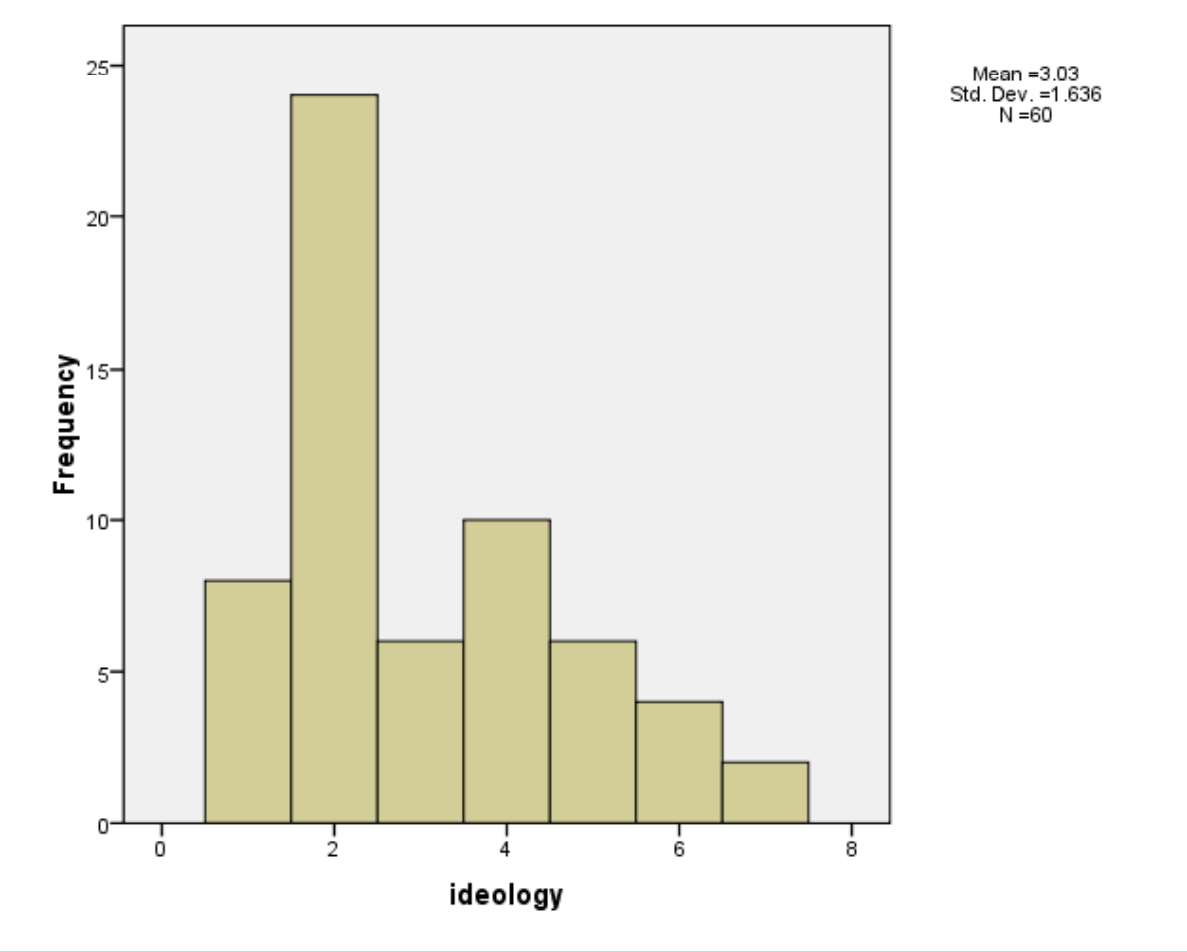
Example: Student survey (n = 60)
www.stat.ufl.edu/~aa/social/data.html

"political ideology" measured as ordinal variable with 1 = very liberal, 4 = moderate, 7 = very conservative

ideology

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	8	13.3	13.3	13.3
	2	24	40.0	40.0	53.3
	3	6	10.0	10.0	63.3
	4	10	16.7	16.7	80.0
	5	6	10.0	10.0	90.0
	6	4	6.7	6.7	96.7
	7	2	3.3	3.3	100.0
	Total	60	100.0	100.0	

Histogram: Bar graph of frequencies or percentages



Shapes of histograms

- *Bell-shaped* (IQ, SAT, political ideology in all U.S.)
- *Skewed right* (annual income, no. times arrested)
- *Skewed left* (score on easy exam)
- *Bimodal* (polarized opinions)

Ex. GSS data on sex before marriage in Exercise 3.73: always wrong, almost always wrong, wrong only sometimes, not wrong at all

category counts 238, 79, 157, 409

Stem-and-leaf plot

Example: Exam scores (*n* = 40 students)

Stem	Leaf
3	6
4	
5	37
6	235899
7	011346778999
8	00111233568889
9	02238

2.Numerical descriptions

Let *y* denote a quantitative variable, with observations y_1 , y_2 , y_3 , … , y_n

a. Describing the *center*

Median: Middle measurement of ordered sample

Mean: $$\overline{y} = \frac{y_1 + y_2 + ... + y_n}{n} = \frac{\Sigma y_i}{n}$$

Example: Annual per capita carbon dioxide emissions (metric tons) for $n = 8$ largest nations in population size

Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, U.S. 20.1

Ordered sample:

Median =

Mean $\bar{y}$ =

Example: Annual per capita carbon dioxide emissions (metric tons) for $n = 8$ largest nations in population size

Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, U.S. 20.1

Ordered sample: 0.3, 0.7, 1.2, 1.4, 1.8, 2.3, 9.9, 20.1

Median =

Mean $\bar{y}$ =

Example: Annual per capita carbon dioxide emissions (metric tons) for $n = 8$ largest nations in population size

Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, U.S. 20.1

Ordered sample: 0.3, 0.7, 1.2, 1.4, 1.8, 2.3, 9.9, 20.1

Median = (1.4 + 1.8)/2 = 1.6

Mean $\bar{y}$ = (0.3 + 0.7 + 1.2 + …+ 20.1)/8 = 4.7

Properties of mean and median

- For symmetric distributions, mean = median
- For skewed distributions, mean is drawn in direction of longer tail, relative to median
- Mean valid for interval scales, median for interval or ordinal scales
- Mean sensitive to “outliers” (median often preferred for highly skewed distributions)
- When distribution symmetric or mildly skewed or discrete with few values, mean preferred because uses numerical values of observations

Examples:

- NY Yankees baseball team in 2006

  mean salary = $7.0 million

  median salary = $2.9 million

  How possible? Direction of skew?

- Give an example for which you would expect

  mean < median

b. Describing *variability*

Range: Difference between largest and smallest observations

(but highly sensitive to outliers, insensitive to shape)

Standard deviation: A “typical” distance from the mean

The *deviation* of observation *i* from the mean is

$$y_i - \bar{y}$$

The ***variance*** of the *n* observations is

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1} = \frac{(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}{n-1}$$

The **standard deviation** *s* is the square root of the variance,

$$s = \sqrt{s^2}$$

Example: Political ideology

- For those in the student sample who attend religious services at least once a week (*n* = 9 of the 60),
- *y* = 2, 3, 7, 5, 6, 7, 5, 6, 4

$$\bar{y} = 5.0,$$

$$s^2 = \frac{(2-5)^2 + (3-5)^2 + \ldots + (4-5)^2}{9-1} = \frac{24}{8} = 3.0$$

$$s = \sqrt{3.0} = 1.7$$

For entire sample (n = 60), mean = 3.0, standard deviation = 1.6, tends to have similar variability but be more liberal

- **Properties of the standard deviation:**
  - $s \geq 0$, and only equals 0 if all observations are equal
  - s increases with the amount of variation around the mean
  - Division by $n - 1$ (not n) is due to technical reasons (later)
  - s depends on the units of the data (e.g. measure euro vs $)
  - Like mean, affected by outliers

- *Empirical rule*: If distribution approx. bell-shaped,
  - about 68% of data within 1 std. dev. of mean
  - about 95% of data within 2 std. dev. of mean
  - all or nearly all data within 3 std. dev. of mean

Example: SAT with mean = 500, s = 100

(sketch picture summarizing data)

Example: y = number of close friends you have

Recent GSS data has mean 7, s = 11

Probably highly skewed: right or left?

Empirical rule fails; in fact, median = 5, mode=4

Example: y = selling price of home in Syracuse, NY.

If mean = $130,000, which is realistic?

$s=0$, $s=1000$, $s=50{,}000$, $s=1{,}000{,}000$

c. Measures of *position*

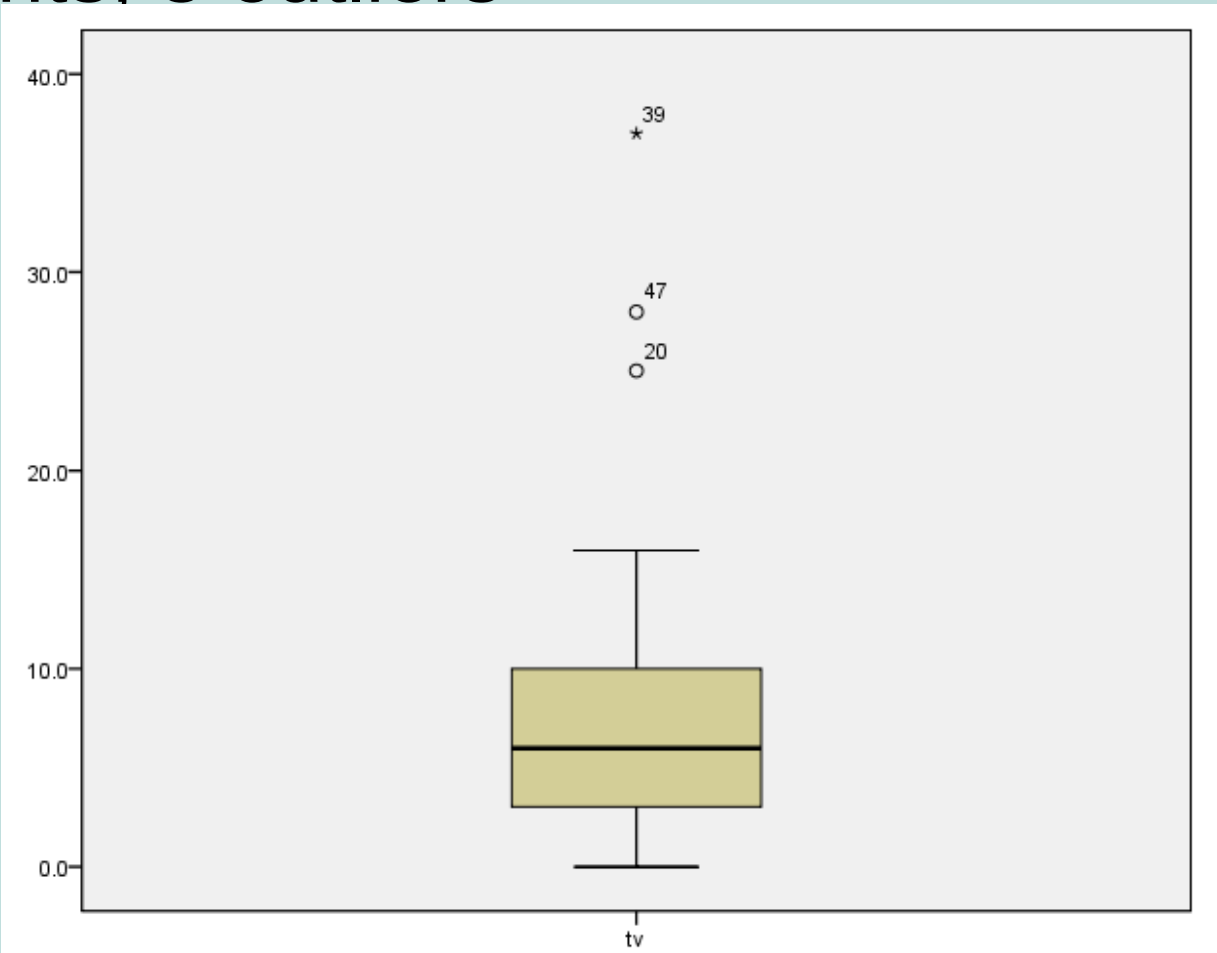
pth *percentile*: *p* percent of observations below it, (100 - *p*)% above it.

- *p* = 50: *median*
- *p* = 25: *lower quartile* (LQ)
- *p* = 75: *upper quartile* (UQ)

- *Interquartile range* IQR = UQ - LQ

Quartiles portrayed graphically by *box plots* (John Tukey 1977)

Example: weekly TV watching for *n*=60 students, 3 outliers



Box plots have box from LQ to UQ, with median marked. They portray a *five-number summary* of the data:

Minimum, LQ, Median, UQ, Maximum

with outliers identified separately

Outlier = observation falling

below LQ – 1.5(IQR)

or above UQ + 1.5(IQR)

Ex. If LQ = 2, UQ = 10, then IQR = 8 and outliers above 10 + 1.5(8) = 22

Bivariate description

- Usually we want to study *associations* between two or more variables (e.g., how does number of close friends depend on gender, income, education, age, working status, rural/urban, religiosity…)
- Response variable: the outcome variable
- Explanatory variable: defines groups to compare

Ex.: number of close friends is a response variable, while gender, income, … are explanatory variables

Response = “dependent variable”

Explanatory = “independent variable”

Summarizing associations:

- Categorical var's: show data using *contingency tables*
- Quantitative var's: show data using *scatterplots*
- Mixture of categorical var. and quantitative var. (e.g., number of close friends and gender) can give numerical summaries (mean, standard deviation) or side-by-side box plots for the groups

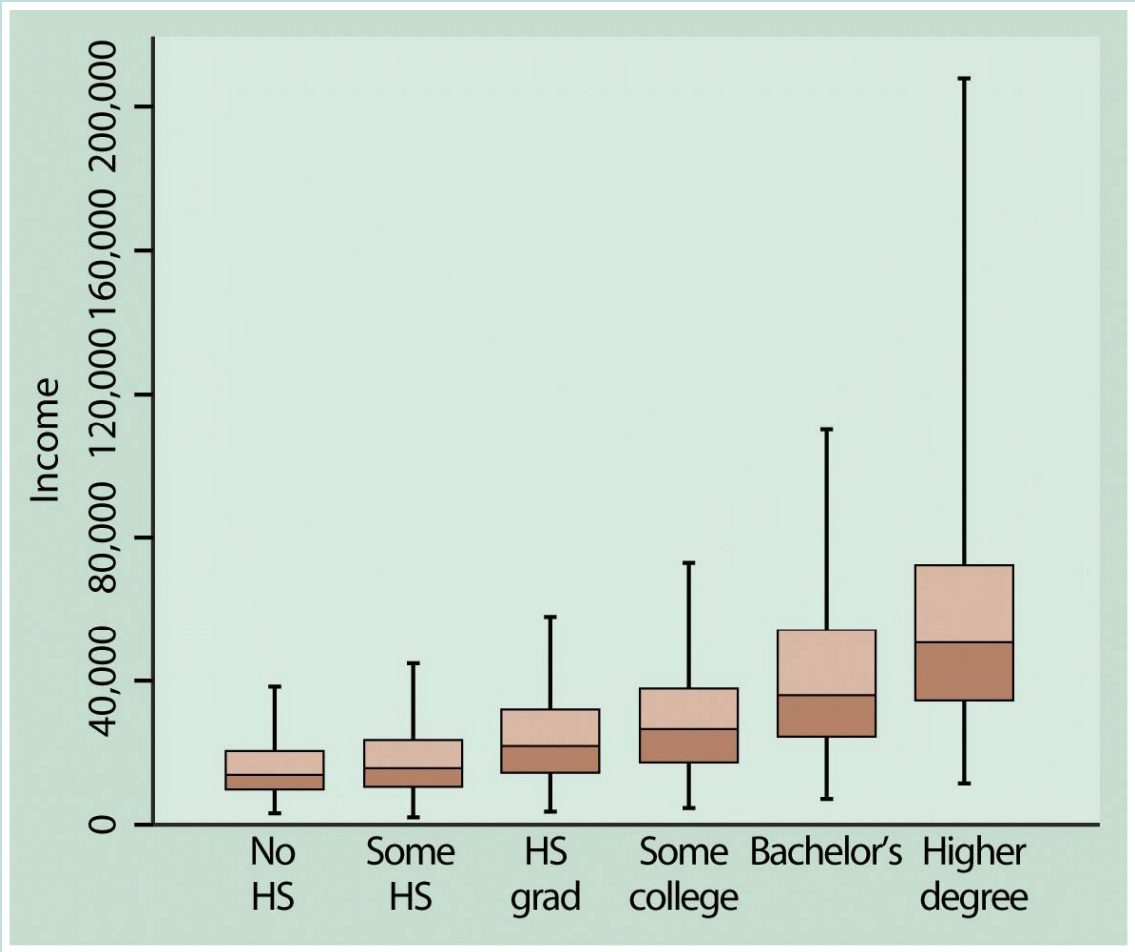
- ***Ex***. General Social Survey (GSS) data

  Men: mean = 7.0, s = 8.4

  Women: mean = 5.9, s = 6.0

Shape? Inference questions for later chapters?

Example: Income by highest degree



Contingency Tables

- Cross classifications of categorical variables in which rows (typically) represent categories of explanatory variable and columns represent categories of response variable.

- Numbers in “cells” of the table give the numbers of individuals at the corresponding combination of levels of the two variables

Happiness and Family Income (GSS 2006 data)

	Happiness			
Income	Very	Pretty	Not too	Total
Above Aver.	272	294	49	615
Average	454	835	131	1420
Below Aver.	185	527	208	920
Total	911	1656	388	2955

Can summarize by percentages on response variable (happiness)

Example: Percentage “very happy” is

44% for above aver. income (272/615 = 0.44)

32% for average income (454/1420 = 0.32)

20% for below average income

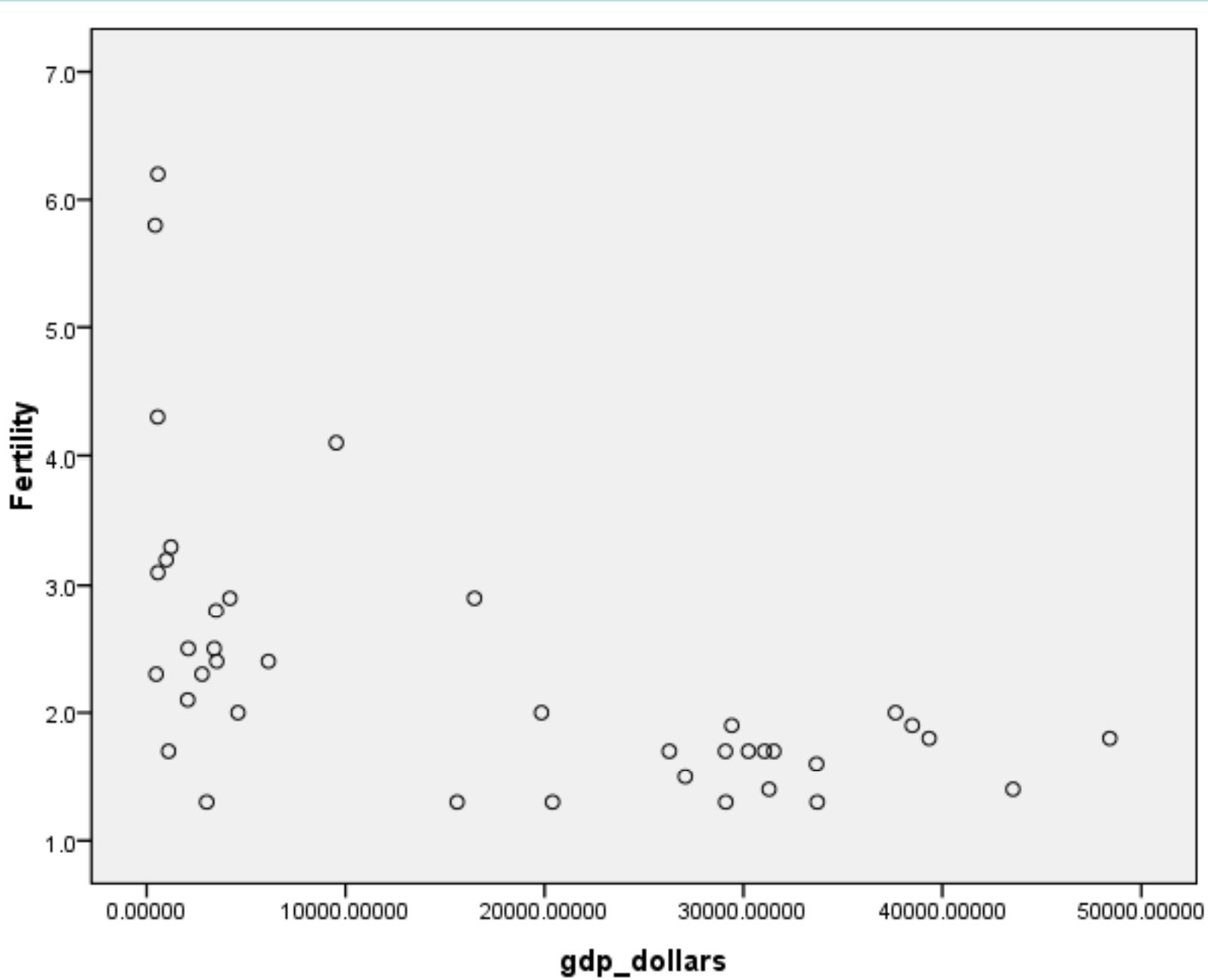
	Happiness			
Income	Very	Pretty	Not too	Total
Above	272 (44%)	294 (48%)	49 (8%)	615
Average	454 (32%)	835 (59%)	131 (9%)	1420
Below	185 (20%)	527 (57%)	208 (23%)	920

Inference questions for later chapters? (i.e., what can we conclude about the corresponding population?)

Scatterplots (for quantitative variables) plot response variable on vertical axis, explanatory variable on horizontal axis

Example*:* Table 9.13 (p. 294) shows UN data for several nations on many variables, including fertility (births per woman), contraceptive use, literacy, female economic activity, per capita gross domestic product (GDP), cell-phone use, CO2 emissions

Data available at http://www.stat.ufl.edu/~aa/social/data.html


Fertility
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.00000
10000.00000
20000.00000
30000.00000
40000.00000
50000.00000
gdp_dollars

Example: Survey in Alachua County, Florida, on predictors of mental health

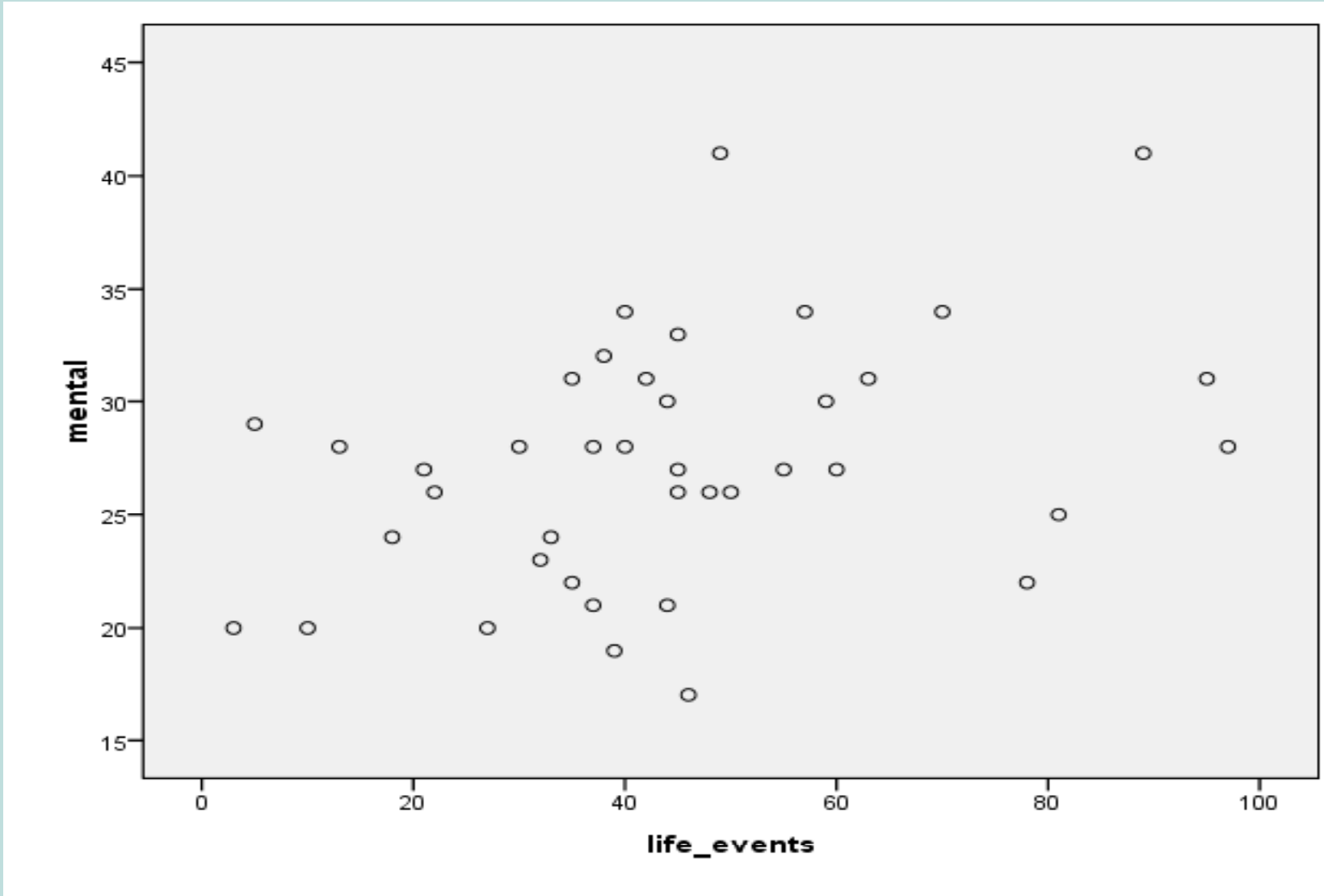
(data for $n = 40$ on p. 327 of text and at www.stat.ufl.edu/~aa/social/data.html)

y = measure of mental impairment (incorporates various dimensions of psychiatric symptoms, including aspects of depression and anxiety)

(min = 17, max = 41, mean = 27, $s = 5$)

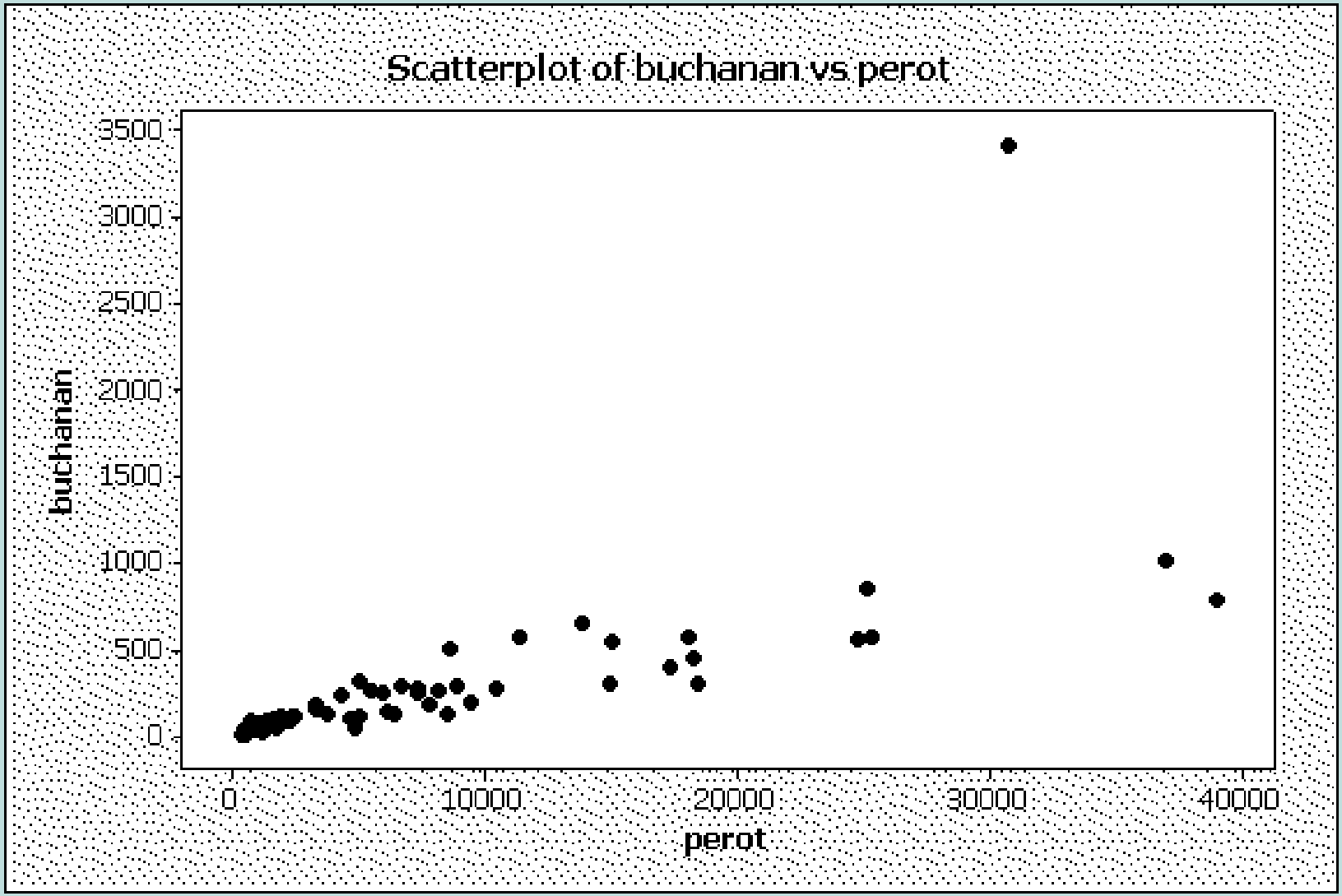
x = life events score (events range from severe personal disruptions such as death in family, extramarital affair, to less severe events such as new job, birth of child, moving)

(min = 3, max = 97, mean = 44, $s = 23$)


mental
45
40
35
30
25
20
15
0
20
40
60
80
100
life_events

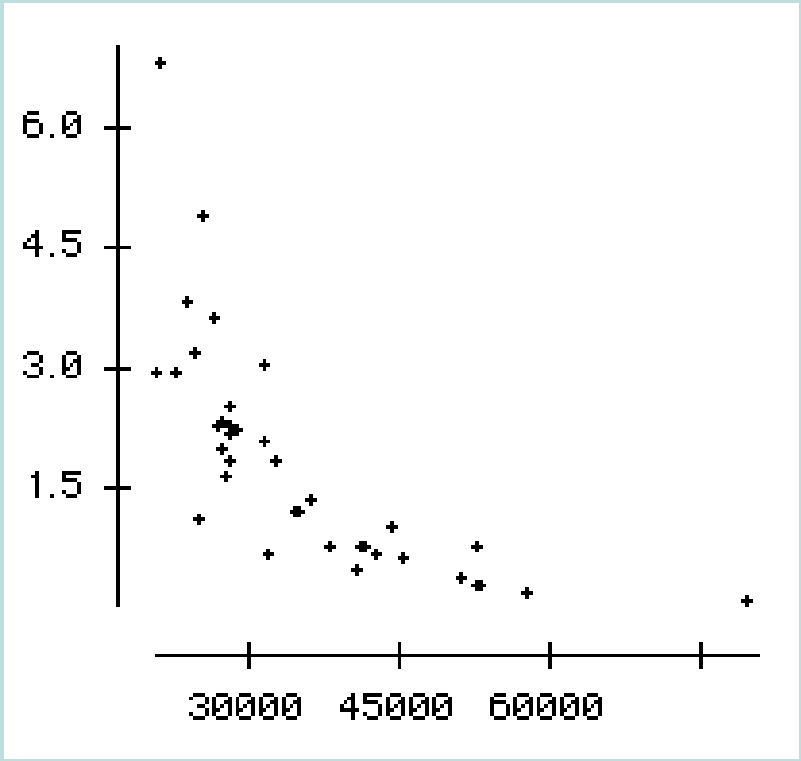
Bivariate data from 2000 Presidential election

Butterfly ballot, Palm Beach County, FL, text p.290



Example: The Massachusetts Lottery
(data for 37 communities, from Ken Stanley)

% income spent on lottery



Per capita income

Correlation describes strength of association

- Falls between -1 and +1, with sign indicating direction of association (formula later in Chapter 9)

The larger the correlation in absolute value, the stronger the association (in terms of a straight line trend)

Examples: (positive or negative, how strong?)

Mental impairment and life events, correlation =

GDP and fertility, correlation =

GDP and percent using Internet, correlation =

Correlation describes strength of association

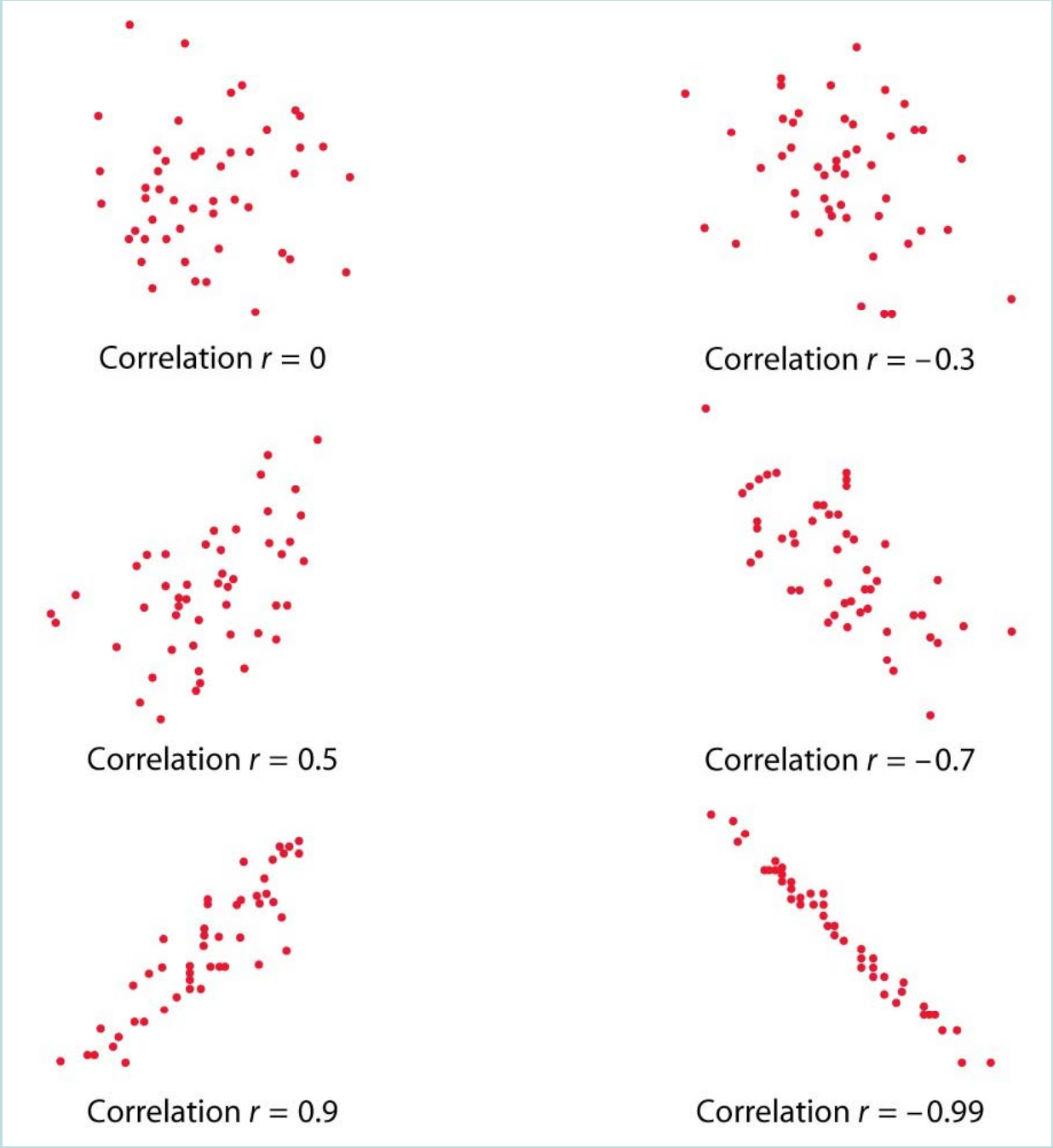
- Falls between -1 and +1, with sign indicating direction of association

Examples: (positive or negative, how strong?)

Mental impairment and life events, correlation = 0.37

GDP and fertility, correlation = -0.56

GDP and percent using Internet, correlation = 0.89


Correlation r = 0
Correlation r = −0.3
Correlation r = 0.5
Correlation r = −0.7
Correlation r = 0.9
Correlation r = −0.99

Regression analysis gives line predicting *y* using *x*

Example:

y = mental impairment, *x* = life events

Predicted y = 23.3 + 0.09x

e.g., at *x* = 0, predicted *y* =

at *x* = 100, predicted *y* =

Regression analysis gives line predicting *y* using *x*

Example:

y = mental impairment, *x* = life events

Predicted y = 23.3 + 0.09x

e.g., at *x* = 0, predicted *y* = 23.3

at *x* = 100, predicted *y* = 23.3 + 0.09(100) = 32.3

Inference questions for later chapters?

(i.e., what can we conclude about the population?)

Example: *y* = college GPA, *x* = high school GPA for student survey

What is the correlation?

What is the estimated regression equation?

We'll see later in course the formulas that software uses to find the correlation and the "best fitting" regression equation

Sample statistics / Population parameters

- We distinguish between summaries of *samples* (**statistics**) and summaries of *populations* (**parameters**).

- Common to denote statistics by Roman letters, parameters by Greek letters:

  Population mean = μ, standard deviation = σ, proportion π are parameters.

  In practice, parameter values unknown, we make inferences about their values using sample statistics.

- The sample mean $\bar{y}$ estimates
  the population mean μ (quantitative variable)

- The sample standard deviation *s* estimates
  the population standard deviation σ (quantitative variable)

- A sample proportion *p* estimates
  a population proportion π (categorical variable)