Chapter 4

Scatterplots and Correlation

BPS - 5th Ed. Chapter 4 1

Explanatory and Response Variables

- Interested in studying the relationship between two variables by measuring both variables on the same individuals.
  - a *response variable* measures an outcome of a study
  - an *explanatory variable* explains or influences changes in a response variable
  - sometimes there is no distinction

BPS - 5th Ed. Chapter 4 2

Question

In a study to determine whether surgery or chemotherapy results in higher survival rates for a certain type of cancer, whether or not the patient survived is one variable, and whether they received surgery or chemotherapy is the other. Which is the explanatory variable and which is the response variable?

BPS - 5th Ed. Chapter 4 3

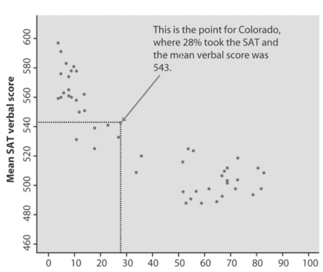
Scatterplot

- Graphs the relationship between two quantitative (numerical) variables measured on the same individuals.
- If a distinction exists, plot the explanatory variable on the horizontal (x) axis and plot the response variable on the vertical (y) axis.

BPS - 5th Ed. Chapter 4 4

Scatterplot

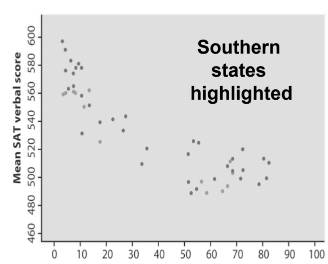
Relationship between mean SAT verbal score and percent of high school grads taking SAT



BPS - 5th Ed. Chapter 4 5

Scatterplot

To add a *categorical variable*, use a different plot color or symbol for each category



BPS - 5th Ed. Chapter 4 6

Scatterplot

- Look for *overall pattern* and *deviations* from this pattern
- Describe pattern by *form*, *direction*, and *strength* of the relationship
- Look for *outliers*

BPS - 5th Ed. Chapter 4 7

Linear Relationship

Some relationships are such that the points of a scatterplot tend to fall along a straight line -- linear relationship

BPS - 5th Ed. Chapter 4 8

Direction

- Positive association
  - above-average values of one variable tend to accompany above-average values of the other variable, and below-average values tend to occur together
- Negative association
  - above-average values of one variable tend to accompany below-average values of the other variable, and vice versa

BPS - 5th Ed. Chapter 4 9

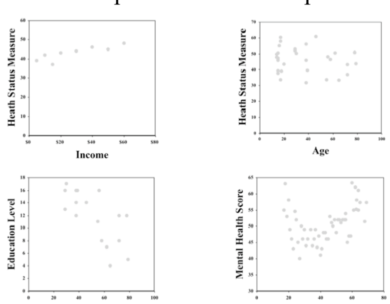
Examples

From a scatterplot of college students, there is a ***positive association*** between verbal SAT score and GPA.

For used cars, there is a ***negative association*** between the age of the car and the selling price.

BPS - 5th Ed. Chapter 4 10

Examples of Relationships



BPS - 5th Ed. Chapter 4 11

Measuring Strength & Direction of a <u>Linear</u> Relationship

- How closely does a non-horizontal straight line fit the points of a scatterplot?
- The correlation coefficient (often referred to as just ***correlation***): **r**
  - measure of the *strength* of the relationship: the stronger the relationship, the larger the magnitude of r.
  - measure of the *direction* of the relationship: positive r indicates a positive relationship, negative r indicates a negative relationship.

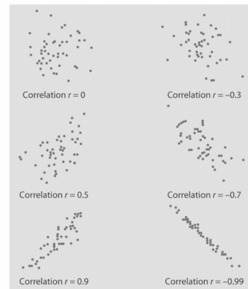
BPS - 5th Ed. Chapter 4 12

Correlation Coefficient

- special values for r :
  - a perfect positive linear relationship would have r = +1
  - a perfect negative linear relationship would have r = -1
  - if there is no *linear* relationship, or if the scatterplot points are best fit by a horizontal line, then r = 0
  - *Note: r must be between -1 and +1, inclusive*
- both variables must be quantitative; no distinction between response and explanatory variables
- r has no units; does not change when measurement units are changed (ex: ft. or in.)

BPS - 5th Ed. Chapter 4 13

Examples of Correlations



BPS - 5th Ed. Chapter 4 14

Examples of Correlations

- Husband's versus Wife's ages
  - r = .94
- Husband's versus Wife's heights
  - r = .36
- Professional Golfer's Putting Success: Distance of putt in feet versus percent success
  - r = -.94

BPS - 5th Ed. Chapter 4 15

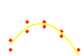
Not all Relationships are Linear Miles per Gallon versus Speed

- Linear relationship?
- Correlation is close to zero.

BPS - 5th Ed. Chapter 4 16

Not all Relationships are Linear Miles per Gallon versus Speed

- Curved relationship.
- Correlation is misleading.



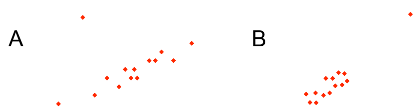
BPS - 5th Ed. Chapter 4 17

Problems with Correlations

- Outliers can inflate or deflate correlations (see next slide)
- Groups combined inappropriately may mask relationships (a third variable)
  - groups may have different relationships when separated

BPS - 5th Ed. Chapter 4 18

Outliers and Correlation



For each scatterplot above, how does the outlier affect the correlation?

A: outlier decreases the correlation
B: outlier increases the correlation

Correlation Calculation

- Suppose we have data on variables X and Y for n individuals:

  $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$

- Each variable has a mean and std dev:

  $(\bar{x}, s_x)$ and $(\bar{y}, s_y)$ (see ch. 2 for s)

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Case Study

Per Capita Gross Domestic Product and Average Life Expectancy for Countries in Western Europe

Case Study

Country	Per Capita GDP (*x*)	Life Expectancy (*y*)
Austria	21.4	77.48
Belgium	23.2	77.53
Finland	20.0	77.32
France	22.7	78.63
Germany	20.8	77.17
Ireland	18.6	76.39
Italy	21.5	78.51
Netherlands	22.0	78.15
Switzerland	23.8	78.99
United Kingdom	21.2	77.37

Case Study

x	y	$(x_i - \bar{x})/s_x$	$(y_i - \bar{y})/s_y$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
21.4	77.48	-0.078	-0.345	0.027
23.2	77.53	1.097	-0.282	-0.309
20.0	77.32	-0.992	-0.546	0.542
22.7	78.63	0.770	1.102	0.849
20.8	77.17	-0.470	-0.735	0.345
18.6	76.39	-1.906	-1.716	3.271
21.5	78.51	-0.013	0.951	-0.012
22.0	78.15	0.313	0.498	0.156
23.8	78.99	1.489	1.555	2.315
21.2	77.37	-0.209	-0.483	0.101
$\bar{x}$ = 21.52, s_x = 1.532	$\bar{y}$ = 77.754, s_y = 0.795			sum = 7.285

Case Study

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \left(\frac{1}{10-1}\right)(7.285)$$

$$= 0.809$$