

CHAPTER 7

Scatterplots, Association, and Correlation

Examining Relationships

- Relationship between 2 or more variables
 - Ex. Height and Weight
 - Ex. Alcohol and Body Temperature
 - Ex. SAT Verbal Score and SAT Math Score
 - Ex. High School GPA and College GPA

Two Types of Variables

- Response Variable
 - Measures the outcome of the study
- Explanatory Variable
 - Used to explain the response variable
- Example
 - Body Temperature = Response Variable
 - Alcohol = Explanatory Variable

Two Types of Variables

- Does not imply causation
 - Explanatory variable does not cause the response variable
- Sometimes there is no true response or explanatory variables
 - Ex. Height and Weight
 - Ex. SAT verbal and SAT math scores

Graphing Two Variables Scatterplot

- Plot of response variable vs. explanatory variable
 - Explanatory variable is plotted on the horizontal axis/x-axis
 - Response variable is plotted on the vertical axis/y-axis
- If there is no true response/explanatory variables, you can plot the variables on either axis

Interpreting Scatterplots

- Scatterplots allows us to observe patterns, trends, and relationships
- When observing a scatterplot always look at the following
 - Overall Pattern
 - Form
 - Direction
 - Strength
 - Deviations from the pattern
 - outliers

Interpreting Scatterplots

- Form
 - Is the plot linear? Is the plot curved? Is there a distinct pattern in the plot?
- Strength
 - Does the plot follow the form very closely? Or is there a lot of variation?

Interpreting Scatterplots

- Direction
 - Is the plot increasing? Is the plot decreasing?
 - Positively Associated
 - Higher (smaller) values in one variable are associated with higher (smaller) values in the other variable
 - Negatively Associated
 - Higher (smaller) values in one variable are associated with smaller (higher) values in the other variable

Example-Scatterplot

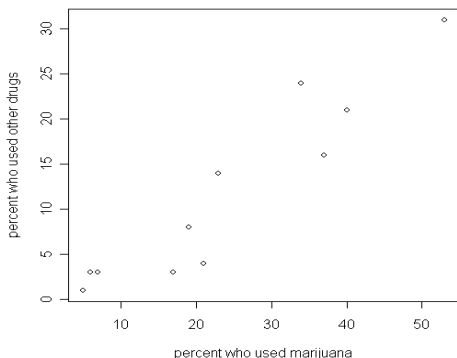
- The following survey was conducted in the United States and 10 Western European countries to determine the percentage of teenagers who have used marijuana and other drugs

Example-Scatterplot

country	percent who have used	
	marijuana	other drugs
Czech Republic	21	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
North Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
United States	34	24

Example-Scatterplot

teenage drug use in 11 countries



Example-Scatterplot

- No true explanatory or response variable
- Form
 - linear
- Strength
 - moderately strong
- Direction
 - positive
- no outliers

Correlation

- Measures the strength of a **linear** relationship between two variables
 - Denoted by r

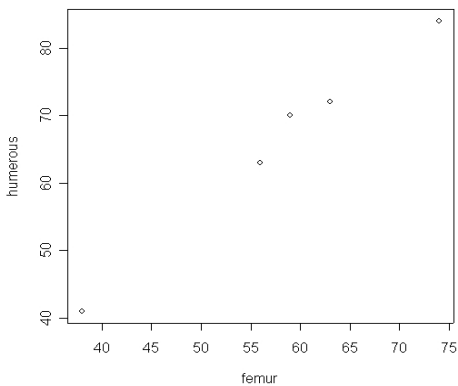
$$r = \frac{1}{n-1} \left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \right)$$

Correlation

- Steps to find correlation
 - Calculate the mean of x and y
 - Calculate the standard deviation for x and y
 - Calculate $\sum (x - \bar{x})(y - \bar{y})$
 - Plug all numbers into formula

Correlation

femur vs. humerus



Example-Calculate r

- Femur(x): 38 56 59 63 74
- Humerus(y): 41 63 70 72 84
- Make a table

Example-Calculate r

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
38	41	-20	-25	400	625	500
56	63	-2	-3	4	9	6
59	70	1	4	1	16	4
63	72	5	6	25	36	30
74	84	16	18	256	324	288
290	330	0	0	686	1010	828

Example-Calculate r

- Using the table to calculate mean and standard deviation for each variable

$$\bar{x} = \frac{290}{5} = 58$$

$$s_x = \sqrt{\frac{686}{4}} = 13.1$$

$$\bar{y} = \frac{330}{5} = 66$$

$$s_y = \sqrt{\frac{1010}{4}} = 15.9$$

Example-Calculate r

- Now plug everything into the formula

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$
$$= \frac{828}{(5 - 1)(13.1)(15.9)}$$
$$= 0.994$$

Correlation

- Properties of r
 - r has no units
 - Measures the strength of a LINEAR association between two quantitative variables
 - If the data have a curvilinear relationship, the correlation may not be strong even if the data follow the curve very closely
 - Not resistant to outliers

Correlation

- Properties of r
 - r ranges in values from -1 to 1
 - r = 1 indicates a straight increasing line
 - r = -1 indicates a straight decreasing line
 - r = 0 indicates no LINEAR relationship
 - As r moves away from 0, the linear relationship between variables is stronger
 - Changing the scale of x or y will not change the value of r

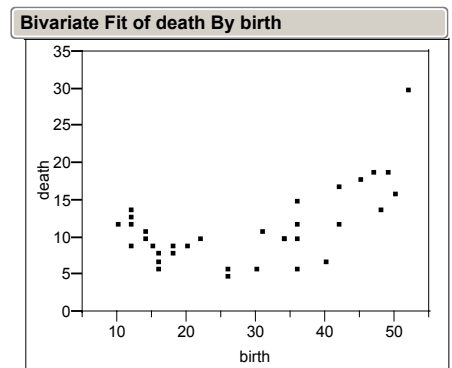
Correlation

- Properties of r
 - Strong correlation between two variables DOES NOT MEAN that the explanatory variable causes the response variable.
 - Strong linear relationship is NOT PROOF of a causal relationship!

Reading JMP Output

- The following is some output from JMP where I compared the birth rate to the death rate of several countries around the world
- explanatory variable X = birth rate
response variable Y = death rate

Reading JMP Output



Reading JMP Output

Summary of Fit

RSquare	0.318097
RSquare Adj	0.296788
Root Mean Square Error	4.232017
Mean of Response	11.47059
Observations (or Sum Wgts)	34

Reading JMP Output

- $RSquare = r^2$
- $r = \sqrt{RSquare} = \sqrt{0.318097} = 0.564$
- We know this is positive because the scatterplot has a positive direction
- The Mean of the Response is the mean of the y 's or \bar{y}