Chapter 6 - Scatterplots, Association, and Correlation

February 6, 2015

In chapter 6-8, we look at ways to compare the relationship of 2 quantitative variables. First we will look at a graphical representation, and then we will talk about some numerical calculations.

6.1 Scatterplots

Scatterplot: Plot that shows the relationship between 2 quantitative variables. Values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical. Data for each individual is plotted as a point. Typically numerical values for each variable will be positive, so this is like plotting points in the first quadrant from Algebra. Usually, the scales on the axis are adjusted to better represent the data. Usually, the origin (0,0) is not shown.

Explanatory Variable (Predictor Variable): Variable that explains or causes changes in the response variable. This variable is plotted on the horizontal axis.

Response Variable: Variable of interest or that measures an outcome of a study. This variable is plotted on the vertical axis.

If the roles of the variable are not clear, then which variable is placed on which axis is not important.

Example: p. 170: 12

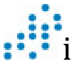
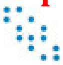
a) Long-distance calls: time (minutes), cost

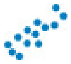
b) Lightning strikes: distance from lightning, time delay of the thunder

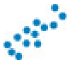
c) A streetlight: its apparent brightness, your distance from it

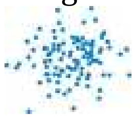
d) Cars: weight of car, age of owner

Look for the overall **pattern** in the scatterplot.

1. **Direction** - A scatterplot that has points that rise from the lower left to the upper right  is called **positive**. A plot that has points that fall from the upper left to the lower right  is called **negative**.

2. **Form** - This is the "Shape" of the points. Points might seem to follow a line (**Linear**)  or some other curve or . We will be looking at linear forms since these tend to be the most common and most useful in Statistics.

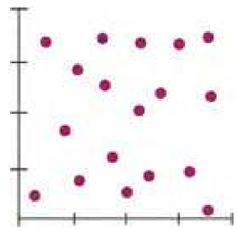
3. **Strength** - How spread out the points are from the form. Data is considered **strong** if it is clustered around the form whether straight  or curved .

   The other extreme there is no recognizable shape 

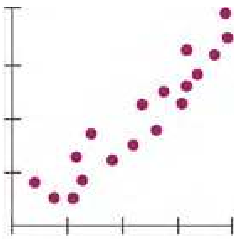
4. **Unusual Features** - Look for any deviations from the pattern. These could be **outliers** (Individual points away from the pattern) or **subgroups** (Clusters of points away from the pattern).

Example: p. 170: 14

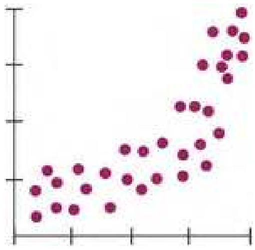
1.



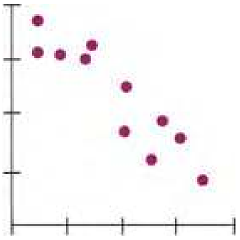
2.



3.



4.



a) little or no association?

b) a negative association?

c) a linear association?

d) a moderately strong association?

e) a very strong association?

6.2 Correlation

Correlation (Coefficient), *r*, measures the strength and gives the direction of the linear association between two quantitative variables.

In order to use correlation to describe an association the following conditions must be met.

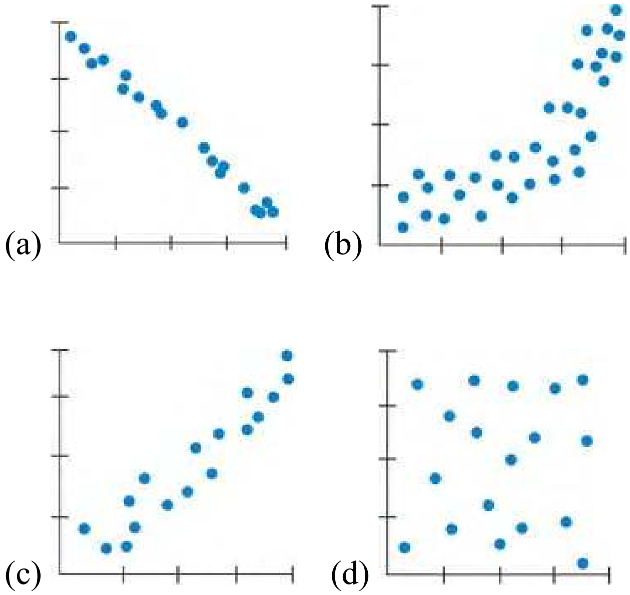
1. Variables must be quantitative.
2. Association must be linear (correlation only measures the strength of a linear association.)
3. Watch out for outliers. (Report correlation with and without outliers.)

$$r = \frac{\sum z_x z_y}{n-1} = \frac{\sum \left(x-\overline{x}\right)\left(y-\overline{y}\right)}{\left(n-1\right)s_x s_y}$$

Correlation Properties

- Sign of the Correlation gives the direction of the association.
- Correlation takes on values between and including -1 to 1. $\left(-1 \le r \le 1\right)$. If Correlation is -1 or 1 than that means data falls exactly on a line. If Correlation is close to -1 or 1 then the association is strong and points will be close to the line. If the Correlation is close to 0 then the points are spread out and there will appear to be no linear association.
- Correlation of x with y is the same as the correlation of y with x.
- Correlation has no units
- Correlation is not affected by changing the center or scale of either variable.
- Correlation measures only the strength of linear associations. (Variables can be strongly associated but if the association is not linear the correlation will be small.)
- Corrrelation is sensitive to outliers. (It is based the z-score which uses the Mean and Standard Deviation.)

p. 172, #**20**. Here are several scatterplots. The calculated correlations are −0.977, −0.021, 0.736, and 0.951. Which is which?



p. 174, #42. Drug abuse A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the following table.

	Percent Who Have Used	
Country	**Marijuana**	**Other Drugs**
Czech Rep.	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No. Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
United States	34	24

a) Create a scatterplot.
b) What is the correlation between the percent of teens who have used marijuana and the percent who have used other drugs?
c) Write a brief description of the association.
d) Do these results confirm that marijuana is a "gateway drug," that is, that marijuana use leads to the use of other drugs? Explain.

What Can Go Wrong? (p. 163-165)

- Do not say “correlation” when you mean “association.

- Do not correlate categorical variables.

- Do not think that correlation shows causation. Just because there is an association between x and y does not mean x causes y.
  Watch out for **Lurking Variables** (A variable in the background other than x and y that simultaneously affects both variables, accounting for the correlation between the two).

- Make sure association is linear

- Do not assume association is linear if the correlation is high.

- Beware of outliers.