Ch. 8: Correlation

The methods described so far (histograms, averages, SD's, etc.) are good for looking at one variable at a time, but we need something else to look at the relationship between two variables.

Ex: Karl Pearson's 1903 study of the heights of 1,078 fathers and their (adult) sons.

A _____ or _____ allows us to interpret a large number of values of _____ variables.

In a scatterplot, each point represents a single pair. In this case, the x-coordinate of the point is the father's height, while the y-coordinate of the point is his son's height. The variables are usually chosen so that we can think of the _____ (or _____) (x) variable as having some influence on the _____ (or _____) (y) variable.

The main mass of points is shaped roughly like a football. This is a common shape for scatterplots. (If each variable's histogram closely follows a normal curve, the joint scatterplot will be football shaped.)

The collection of points slopes up. This is called _____ (taller fathers tend to have taller sons).

If taller fathers tended to have shorter sons, there would be a downward slope to the graph, and we would call this _____.

If a father's height always exactly equaled his son's height, all the points would fall on a straight line (_____). Knowing a father's height would tell us his son's height exactly.

If they were not exactly equal, but were always close, the points would fall in a narrow band around a straight line (_____). Knowing a father's height would give us a very good idea of his son's height.

If we had only _____, the points would fall very broadly around the line.

Ex: What do you think: How much does it help to know a father's height to predict his son's height?

Ex: What type of association would you expect (positive/negative, weak/strong) between the ages of the husbands and wives in a large survey of married couples? Why?

The SD line

The line the points cluster around is called the _____.

The SD line always goes through the point of averages.

When the two variables have a positive association, the slope of the SD line is

$$\text{slope} = \ldots$$

If they have a negative association, the slope of the SD line is

$$\text{slope} = \ldots$$

All points on the line are the same number of SD's away horizontally as vertically from the respective averages.

Suppose we have a scatterplot of two variables and we want to summarize it numerically. What numbers should we use?

1. The average of x. The horizontal center of the point cloud.
2. The average of y. The vertical center of the point cloud. The *point of averages* (x-avg., y-avg.) gives the center of the cloud.
3. The SD of x. The horizontal spread of the cloud.
4. The SD of y. The vertical spread of the cloud.

These still don't describe the relationship of the two variables. We'll measure clustering around a line by

5. The _____.

The correlation coefficient is a unitless number between -1 and 1.

If r is positive, the two variables show _____ association; if r is negative, there is a _____ association.

Numbers close to 1 or -1 show _____ and _____ correlation, respectively. If r *equals* 1 or -1, the two variables have _____ correlation, and are falling exactly on a line. If r is 0, we say there is _____ correlation between the two variables. This means there is no *linear* relationship between the two variables (although it is possible that there is a nonlinear one).

Weak correlations (say, .1 to .5) are common, especially in the social sciences.

Note that there is no direct way to interpret the exact value of the correlation coefficient. If r is 0.6, that doesn't mean that 60% of the values are clustered around the line, or even that it is "twice" as linear as an r of 0.3.

To compute r, convert each variable to standard units. The average of the products is the correlation coefficient.

Ex:

x	y	x in s.u.	y in s.u.	product
1	5			
3	9			
4	7			
5	1			
7	13			