Linear Correlation

Tim Peil's Video's on Using Microsoft Excel to Compute Statistics and Linear Correlation:

1. Go to http://www.mnstate.edu/peil/M102/videos.htm
2. Scroll down to the section titled: "Microsoft Excel for Statistics"
3. Watch the videos: "Internal Statistics Functions"(3:43) and "Line of Best Fit"(4:46)

Definitions:

• We say there is a **correlation** between two variables if one variable is related to the other in some way.

• We say the variables are **linearly correlated** if the correlation between them can be described "well" using a linear function.

• The **linear correlation coefficient** between two variables is a number r between -1 and 1 that describes the degree to which a line can be used to describe the relationship between the two variables.

If $r > 0$ then x and y are *positively correlated.* That is, as x increases, y also increases.

If $r < 0$ then x and y are *negatively correlated.* That is, as x increases, y decreases.

The closer r is to ± 1, the stronger the correlation between x and y is. That is, the better a line "fits" the data.

• The **line of best fit** between two variables is the line that best represents the linear relationship that exists between two variables. It is the line that minimizes the sum of the vertical distance of each data point from the line relating the two variables.

Formulas:

• Given n data points of the form (x, y), The *correlation coefficient* is given by the formula:

$$r = \frac{n \sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2} \cdot \sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

• Given n data points of the form (x, y), The *slope of the line of best fit* is given by the formula:

$$m = \frac{n \sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

and the y-intercept of the line of best fit is given by:

$$b = \frac{\sum y - m\left(\sum x\right)}{n}$$

Example:

Given the following frequency table:

x	y	x^2	y^2	xy
2	4	4	16	8
3	8	9	64	24
4	7	16	49	28
6	10	36	100	60
$\sum x = 15$	$\sum y = 29$	$\sum x^2 = 65$	$\sum y^2 = 229$	$\sum xy = 120$

Notice that $n = 4$

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2} \cdot \sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} = \frac{(4)(120) - (15)(29)}{\sqrt{4(65) - (15)^2} \cdot \sqrt{4(229) - (29)^2}} = \frac{480 - 435}{\sqrt{35}\sqrt{75}} \approx .8783$$

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{4(120) - (15)(29)}{4(65) - 15^2} = \frac{45}{35} \approx 1.2857$$

$$b = \frac{\sum y - m\left(\sum x\right)}{n} \approx \frac{29 - (1.2857)(15)}{4} \approx 2.4286$$

Therefore, the line of best fit for the data in this example is: $y = 1.2857x + 2.4286$.

In this example, the correlation coefficient is .8783, which is fairly close to 1, so the data is fairly linear. (Read page 841-842 in your text and see the table on the top left of page 842 to see that since n is so small, we *cannot* be 95% confident that the data is linearly correlated in this example)