Association between variables, correlation

Two variables are *associated* if certain values of one tend to occur more often with some values of the second than with other values.

If we wish to use changes or values in one variable to explain changes or values in the other, we refer to them as the *explanatory variable*—measuring the cause or explanation (think of x in our usual function notation)—and the *response variable*—measuring the outcome (think of y in our usual use of function and graphing notation).

For two quantitative variables, the standard picture is a *scatterplot*—explanatory variable on the x-axis, response on the y-axis, plot the (x, y) - pairs. [If there isn't an explanatory variable—if we aren't explaining one by the other—either variable can go on either axis].

Patterns in the graph [if there are any] are described by *form* (linear or type of curve—general shape of graph), *direction* (positive—variables increase/decrease together; or negative— when one increases, the other decreases; sometimes there is no fixed direction) and *strength* (how closely packed around the basic pattern are the data points?).

Correlation is a measurement of the amount of linear ($y = ax + b$) association between two quantitative variables.

The *correlation coefficient* r is given by $r = \dfrac{1}{n-1}\sum\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right)$—but we will calculate using the "Linreg" command on our calculators, or using computer software.

The formula says:

1. Convert all values to Standard units (subtract the mean, divide by the standard deviation of the variable)
2. Multiply each standardized x by the matching standardized y ,
3. Add the results
4. Divide by $n - 1$[to eliminate the effect of more or fewer points being used] .

(That is: we multiply the standardized values of x and y for each point, add them, and divide by $n-1$ to get an average).

The value of r will be between -1 and 1.

- If r is close to 1 , the points are close to making up a line with positive slope
- If r is close to -1 , the points are close to making up a line with negative slope
- If r is close to 0 , the points show no linear association between values of x and values of y [no association, or the association is not linear].

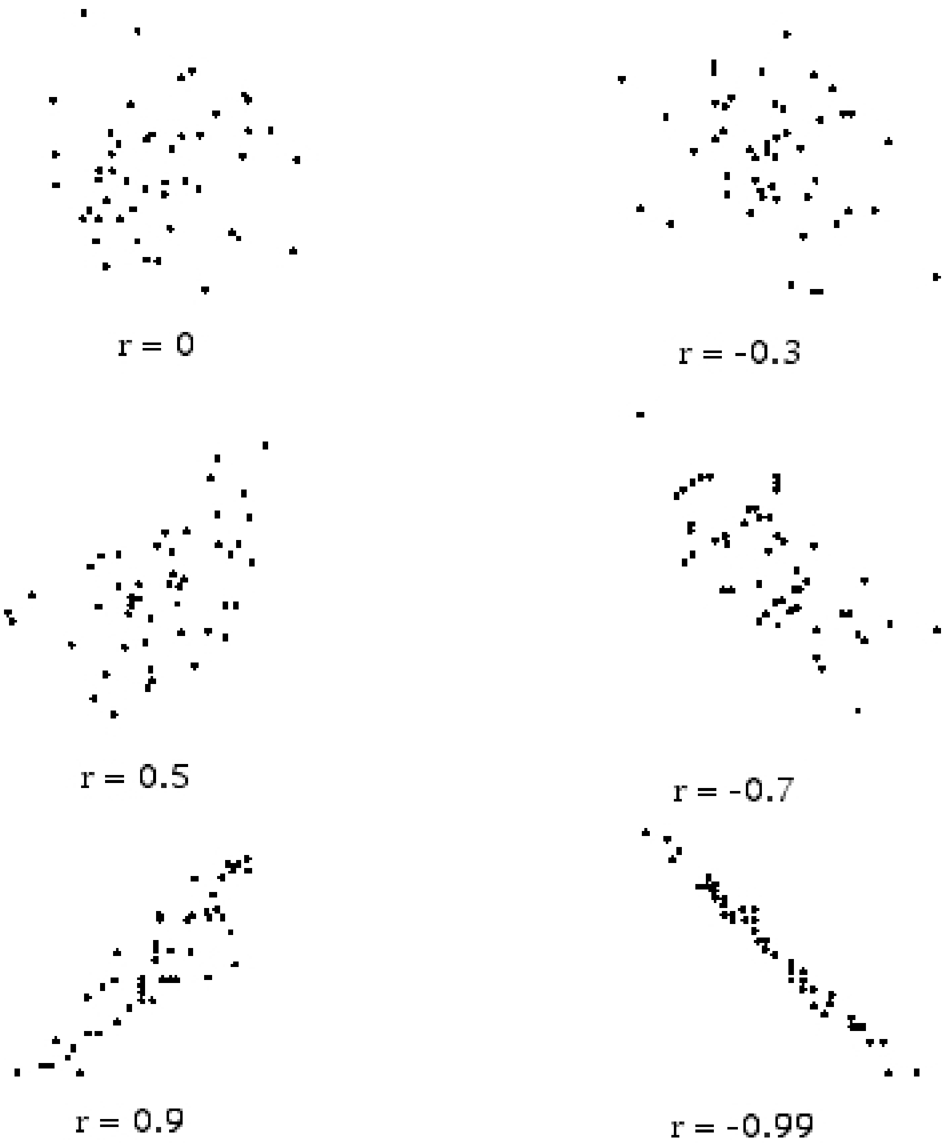
See the examples on the next page showing sets of points with different levels of r . [r closer to 1 or -1 produces tighter clustering around a line. r closer to 0 gives much fuzzier and more diffuse cloud of points. r positive—upward trend in points: r negative— downward trend in points]

What happens:
If large x-values are with large y-values (and small x's with small y's) then the multiplications will give mostly positive numbers (+ times + or $-$ times $-$) and r will be positive, showing a positive association. If large x-values are with small y-values (and small x-values with large y-values) the multiplications will give mostly negative values (+ times $-$ and $-$ times +), so that r will be negative,mshowing negative association. If there is no such pattern, the signs will be mixed, causing cancellation so that r will be close to 0, showing no linear association [r does not measure quadratic, cubic, etc correlation or matching to complicated curves].

In practice, r is not calculated directly from this formula, but with a calculator or computer.

The calculator work will be the same for calculating the regression line equation [coming up]
For TI-83,84: Enter the x-values in one list (using the Stat Edit commands) and then the linreg command [see the instructions for your calculator)—r is displayed as part of the results (TI-84's may require turning this on—see the notes or ask me)
For TI-89— enter data in two columns with the Data Editor, choose calculation, then choose LinReg command
For Casios- use 2-variable data entry, linreg command

Some scatter Plots, showing the corresponding values of the correlation coefficient r


r = 0
r = -0.3
r = 0.5
r = -0.7
r = 0.9
r = -0.99