Correlation

Introduction:

So what have we done so far?

We have compared a *categorical* variable with a *continuous* variable:

For example, t-tests and ANOVA (which we didn't cover) all use continuous variables (e.g., length or height) to compare two (or more, for ANOVA) categorical variables (e.g., males vs. females).

It may seem a bit odd to think about it this way, but it is correct.

We have compared a *categorical* variable with a *categorical* variable:

Contingency tables - for example comparing smoking (yes/no) vs. risk of lung cancer (yes/no).

So now we're ready to compare a *continuous* variable with a *continuous* variable. It turns out there are several ways of doing this. We'll take a look at two:

Correlation - in which we're just interested to see if there's a relationship between two variables.

Regression - in which we're interested in predicting values of one of our variables (or in which one of our variables is obviously *independent*).

For now we're interested only in correlation.

Correlation:

We have two continuous variables, and we want to figure out if there's a relationship between the two. Let's look at the graphs on the following page for example.

Notice that some of the graphs show a pretty strong relationship (positive or negative), while others show a low relationship or even no relationship.

Notice also that each graph has a description using the letter 'r' in the heading.

Notice how r varies:

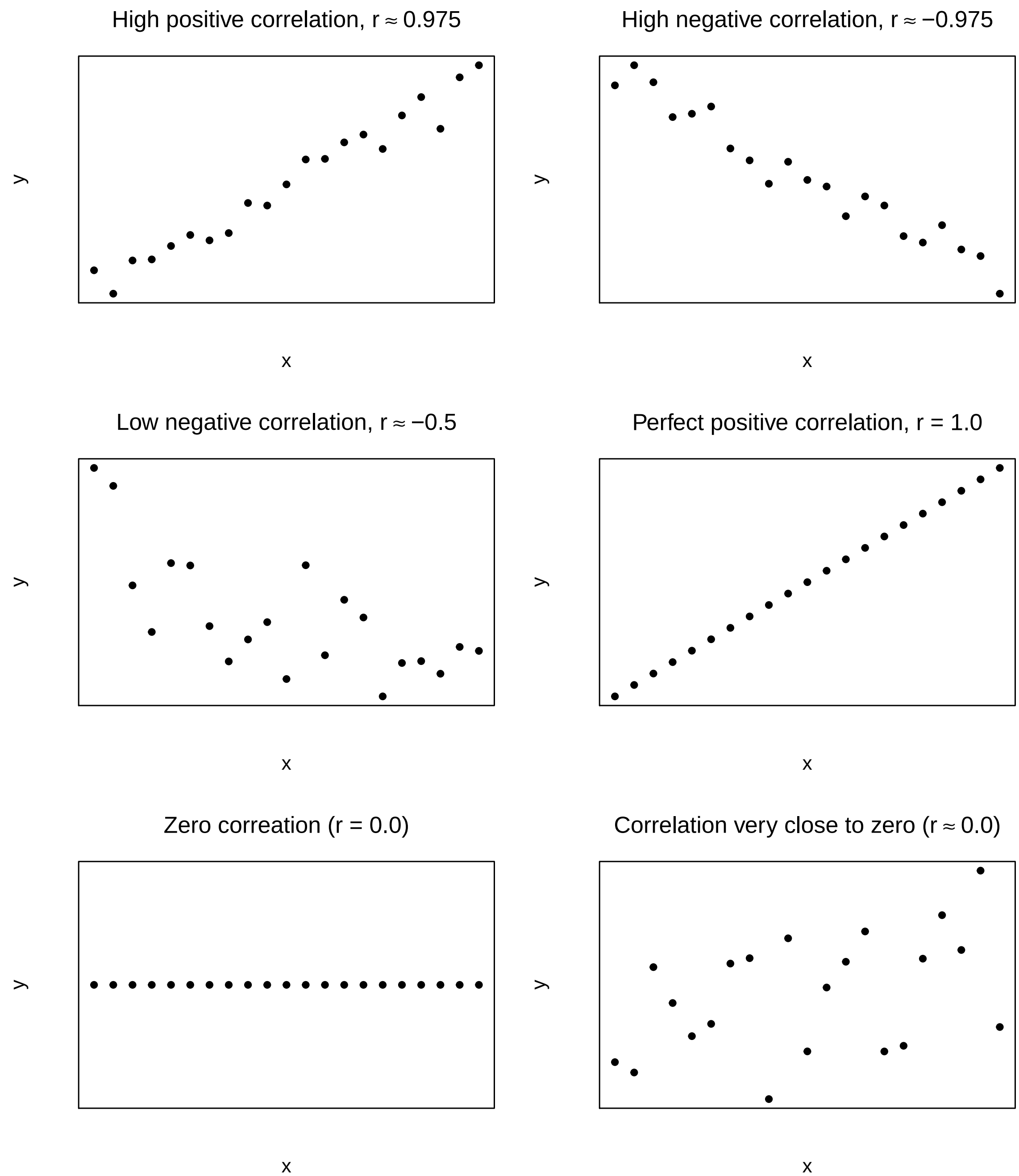
The stronger the relationship, the closer r gets to 1 or -1.

The weaker the relationship, the closer r gets to 0.

A negative relationship is indicated by a (-) value for r.

As usual, r, the sample *correlation coefficient*, is an estimate, in this case of the population correlation coefficient ρ (the Greek letter 'rho'):

$$r \text{ estimates } \rho$$



So how do we calculate r?

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SS_{cp}}{\sqrt{SS_x\ SS_y}}$$

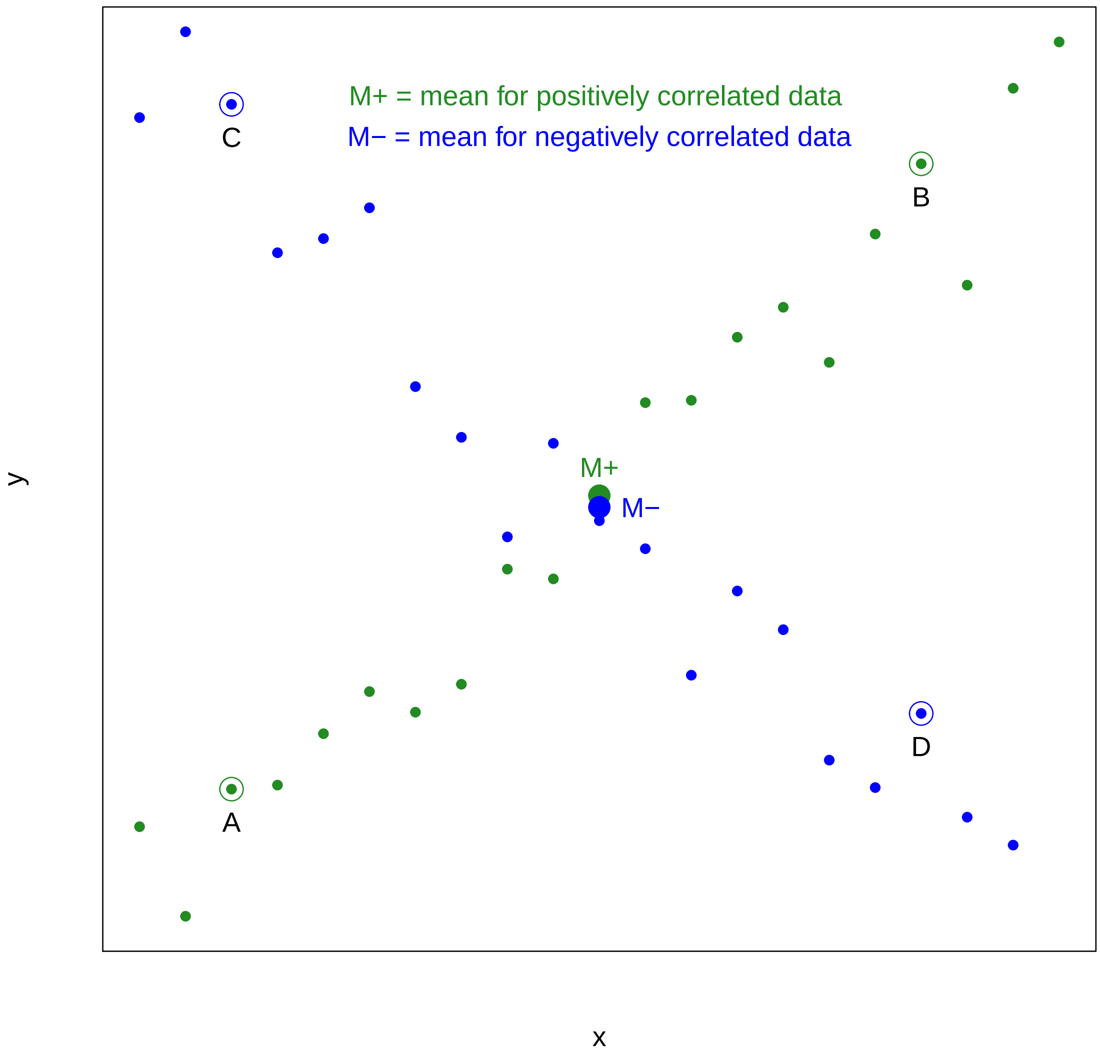
Notice how the numerator is a little like a sum of squares, except that we multiply each $(x_i - \bar{x})$ by each $(y_i - \bar{y})$.

The numerator is sometimes referred to as the sum of **c**ross **p**roducts (and in this class is abbreviated as SS_{cp}).

So how does this equation work?

For the numerator:

Notice that the coordinates for the point $(\bar{x}, \bar{y})$ is somewhere inside our scattering of points (see the graph).



Now suppose that as x increases, y increases (green points), and let's take the point labeled A.

Notice that in this case, $x_i < \bar{x}$ and so $(x_i - \bar{x})$ will be negative.

The same thing holds for y (i.e., $(y_i - \bar{y})$ is negative).

So we are multiplying a (-) number by a (-) number and get a positive number.

For the point B, we're multiplying two positive numbers.

In other words, for points showing a positive relationship we're adding up a bunch of positive numbers in the numerator.

Now suppose that as x increases, y decreases (blue points).

For the point C, notice that $x_i < \bar{x}$ and so $(x_i - \bar{x})$ will be negative.

But now $(y_i - \bar{y})$ is positive, so we're multiplying a (-) number by a (+) number and get a negative number.

The opposite is true for D: we're multiplying a (+) number by a (-) number and again get a negative number.

We're adding up a bunch of negative numbers when we have a negative relationship.

So the numerator changes sign based on the relationship between x and y.

What about the denominator?

Without going into the mathematical details, the numerator makes sure that r stays between -1 and 1. Or in other words, it makes sure that:

$$-1 \leq r \leq 1$$

Notice that if you get a value for r outside these limits *you have made a mistake!*

(If you really want to know why the denominator works the way it does, see your text (one simple example: suppose that each $x_i = y_i$, then it's really easy to see that you get 1)).

So now you know how to calculate the sample correlation coefficient, r. Now what?

r is often used descriptively. Just like $\bar{y}$ is used to describe the mean of a sample, r is often used to describe the correlation between two variables.

However, just because you can describe the correlation, does *not* mean that the relationship is significant.

If you want to see if the relationship is significant, you need to do a hypothesis test.

Testing to see if r shows a significant relationship:

1. Set up your hypotheses as usual:

   $H_0 : \rho = 0$
   $H_1 : \rho \neq 0$

   (Of course we could do $<$ or $>$ for H_1).

2. Decide on α.

3. Calculate r.

4. Calculate $t*$ as follows (yes, we're back to using the t-distribution):

$$t^* = r\sqrt{\frac{n-2}{1-r^2}}$$

5. Compare $|t^*|$ with t_{table} using $n-2$ degrees of freedom (*not* $n-1$).

A one sided test is carried out the same way as usual, however you need to make sure that your data agree with the alternative hypothesis:

For example, if $H_1 : \rho < 0$ then make sure $r < 0$.

So let's do an example. We'll use some data that is built into R, and deals with various measurements of four species of irises (the flower). Since the actual data set is a bit long, we'll only use the first 12 records for *Iris setosa*.

So here are the sepal length and width of 12 flowers:

Sepal length (X)	Sepal weight (Y)
5.1	3.5
4.9	3.0
4.7	3.2
4.6	3.1
5.0	3.6
5.4	3.9
4.6	3.4
5.0	3.4
4.4	2.9
4.9	3.1
5.4	3.7
4.8	3.4

We already know how to calculate the following, so we won't go into the details again:

$\bar{x} = 4.9 \qquad \bar{y} = 3.35 \qquad SS_x = 1.04 \qquad SS_y = 0.99$

But we haven't calculated SS_{cp} before, so here is how to do it:

$$\begin{aligned}
\text{for:}\quad & i = 1: & (5.1 - 4.9)(3.5 - 3.35) &= 0.030 \\
& i = 2: & (4.9 - 4.9)(3.0 - 3.35) &= 0.000 \\
& i = 3: & (4.7 - 4.9)(3.2 - 3.35) &= 0.030 \\
& i = 4: & (4.6 - 4.9)(3.1 - 3.35) &= 0.075 \\
& i = 5: & (5.0 - 4.9)(3.6 - 3.35) &= 0.025 \\
& i = 6: & (5.4 - 4.9)(3.9 - 3.35) &= 0.275 \\
& i = 7: & (4.6 - 4.9)(3.4 - 3.35) &= -0.015 \\
& i = 8: & (5.0 - 4.9)(3.4 - 3.35) &= 0.005 \\
& i = 9: & (4.4 - 4.9)(2.9 - 3.35) &= 0.225 \\
& i = 10: & (4.9 - 4.9)(3.1 - 3.35) &= 0.000 \\
& i = 11: & (5.4 - 4.9)(3.7 - 3.35) &= 0.175 \\
& i = 12: & (4.8 - 4.9)(3.4 - 3.35) &= -0.005 \\
& \textbf{Sum} & &= \mathbf{0.82}
\end{aligned}$$

(Yes, calculating SS_{cp} can be a bit tedious!)

So now let's do the actual hypothesis test:

Set up our hypotheses:

$H_0 : \rho = 0$
$H_1 : \rho > 0 \qquad$ *(why?)*

Pick α. Let's go with $\alpha = 0.05$.

Calculate r:

$$r = \frac{SS_{cp}}{\sqrt{SS_x\, SS_y}} = \frac{0.82}{\sqrt{1.04 \times 0.99}} = \frac{0.82}{1.0147} = 0.8081$$

(Check: $r > 0$, which agrees with H_1, so we continue).

Calculate t^*:

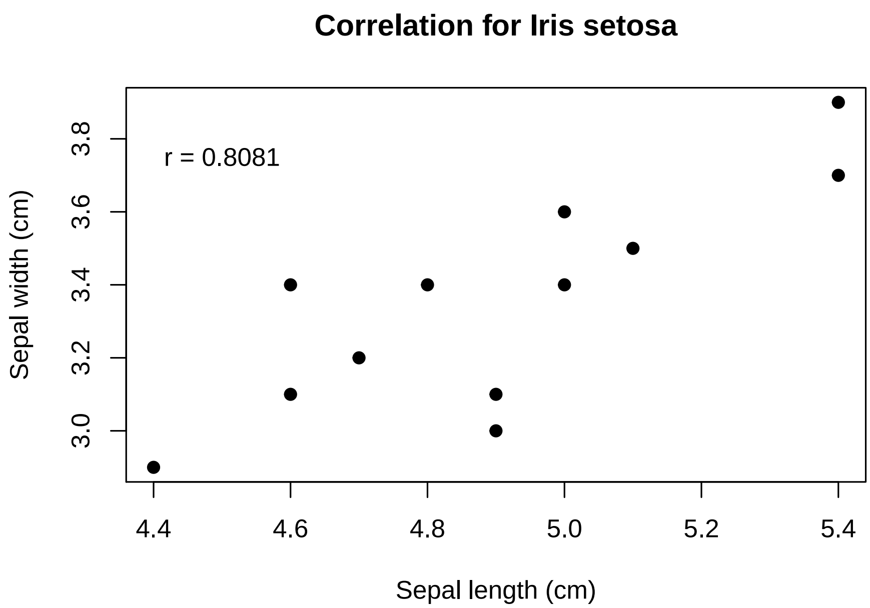
$$t^* = 0.8081\sqrt{\frac{12-2}{1-0.8081^2}} = 4.34$$

And we look up the tabulated value for t using 10 $d.f.$ and find that:

$$t_{0.05,10} = 1.812$$

Since $t^* = 4.66 \geq t_{\text{table}} = 1.812$ we reject H_0 and conclude that weight increases with leaf area.

And here is what it all looks like:



Concluding remarks:

Usually, if r is close to 1 or -1, you'll find that a hypothesis test is significant. *But not always!*

This depends on sample size and other factors:

r might be 0.23 and a hypothesis test highly significant.

r might be 0.95 and a hypothesis test show no significance.

You can not be certain if a correlation is significant without doing a hypothesis test!

r can be sensitive to outliers.

Outliers can pull r in the entirely wrong direction.

(The same is true for regression).

Other than random data, the only assumption we need to worry about in doing a hypothesis test for correlation is that x and y are approximately linear.

In other words, the scatterplot of our data doesn't show any obvious curves in the relationship between x and y.

There is an alternative method (called Spearman's rank correlation) that doesn't care about curves).

This assumption will become *very* important when we do regression.

Finally, just because a hypothesis test is significant does *not* imply cause and effect.

This is particularly true for correlation - a lot of silly things can be correlated which have nothing to do with each other!

True fact: In Europe, the number of storks is highly correlated with the number of babies.

Hopefully everyone realizes that storks don't bring babies.

So what's going on?

To be revealed in class!