Correlation

Comment: notes are adapted from BIOL 214/312.

I. Correlation.

A) Correlation is used when we want to examine the relationship of two continuous variables.

We are not interested in "prediction".

We don't consider one variable "independent" and the other "dependent".

All that we're interested in is:

"Do they vary together?"

Does x go up as y goes up?

or

Does x go down as y goes up?

B) Graphical description:

1) illustrate fig. 19.1, p. 381 on board (include perfect correlation and 0 correlation).

2) Problem - just describing the graphs is subjective, so we describe the correlation using what is called a correlation coefficient.

3) the correlation coefficient is designated by the Greek letter "rho", and is estimated by "*r*". In other words:

r estimates ρ

B) Calculation of "*r*"

1) Here's the formula:

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

2) How does it work?

For the numerator:

a) first notice that the coordinates for $\bar{x}$ and $\bar{y}$ are somewhere inside our points on

the graph (illustrate)

b) if $x_i < \bar{x}$, and $y_i < \bar{y}$, that implies that the point is below and to the left of ($\bar{x}$, $\bar{y}$), and for that value of *i*, the numerator is positive (you're multiplying two negative numbers)

c) similarly, if $x_i > \bar{x}$, and $y_i > \bar{y}$, the point is above and to the right, and the numerator is positive.

- so if all the points pair up below and to the left and above and to the right, *r* will be positive (you're adding up a bunch of positive numbers.

d) now what happens if $x_i < \bar{x}$ and $y_i > \bar{y}$? The point is now above and to the left of ($\bar{x}$, $\bar{y}$), and the numerator is negative (you're multiplying a (+) and (-) number).

e) similarly, if $x_i > \bar{x}$ and $y_i < \bar{y}$, the point is below and to the right, and again the numerator is negative

- if all the points are above to the left and below to the right of ($\bar{x}$, $\bar{y}$), r will be negative.

Also,

- if you have a mix of points, then it depends on where most of the points are, and how far away from ($\bar{x}$, $\bar{y}$) they are.

For the denominator:

a) this is basically a scaling factor. It makes sure that r stays between -1 and 1.

b) This is a little similar to dividing by the standard deviation to get your normal scores (when calculating *z*).

3) Okay, now you have *r*. What do you do next?

- Important - people often use *r*, even if *r* is "not significant". *r* is often used descriptively, sort of like saying the average of something is " $\bar{y}$ ", without saying if $\bar{y}$ means anything other than an "average".

- If you wish to test if the relationship between *x* and *y* is significant, then you carry out a statistical test.

C) Testing for significance in *r*:

a) Set up your hypotheses:

H_0: $\rho = 0$ (this implies that there is no relationship between x and y)
H_1: $\rho \neq 0$ (either positive or negative)

or, of course, H_1: $\rho < 0$, H_1: $\rho > 0$

b) decide on α

c) calculate r

d) calculate t^* (yes, we're back to using t) as follows:

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

Your text does this just a little bit differently, but the formula is actually the same.

e) get your t from the t-tables with α and:

$d.f. = v = n\text{-}2$ (n-2, NOT n-1)!!

f) compare as usual (compare $|t^*|$ to the table t) and make your conclusion.

g) a one sided test is carried out the same way as with a t-test:

- compare the absolute value of t^* with the one sided value from the t-table, and if $t^* \geq t_{table}$, reject.

- just make sure you verify that your data agree with H_1 before you proceed.

D) An example from a different text.

1) A plant physiologist compared the total leaf area with the total dry weight of the plant for 13 plants and got the following results:

Leaf area (X)	Dry weight (Y)	
411	2.00	
550	2.46	$SS_x = 28{,}465.7$
471	2.11	
393	1.89	$SS_y = 0.363708$
427	2.05	
431	2.30	$SS_{cp} = 82.8977$
492	2.46	
371	2.06	(book also gives SS_r which we
470	2.25	haven't discussed yet (we'll
419	2.07	need it for regression))
407	2.17	
489	2.32	
439	2.12	
mean 443.8	2.174	

a) before you go on, what do you think should happen (what kind of alternative hypothesis makes sense)?

b) So let's set up our hypotheses:

H_0: $\rho = 0$ $\quad\quad$ H_1: $\rho > 0$

c) decide on α, so let's pick 0.05

d) calculate *r*:

the book does most of the work for us:

numerator = 82.8977

denominator = square root (28,465.7 x .363708)

So we have:

$$r = \frac{82.8977}{101.75} = .8147$$

(check: since $r > 0$, which agrees with our H_1, we proceed)

e) figure out *t**:

$$t^* = 0.8147\sqrt{\frac{13-2}{1-0.8147^2}} = 4.66$$

f) We look up the tabulated *t* for 11 *d.f.* and $\alpha = 0.05$ and get 1.796

g) We reject H_0 and conclude that leaf area and dry weight increase together.

E) Concluding remarks:

1) there may or not be a direct relationship between *r* and significance:

a) *r* might be .23 and be highly significant

b) *r* might be .95 and not be significant

2) *r* is very sensitive to extreme points (illustrate) (so is regression)

3) while *r* itself doesn't have any assumptions (i.e., you can always calculate *r* (e.g., you can always calculate $\bar{y}$)), the *t*-test based on *r* does. (Normal, random, data). (Caution: obviously, if you use *r* tables, you still have the same assumptions).

4) Because *r* can be sensitive to extreme points, and because sometimes you can't meet the assumptions, there are alternatives:

- Spearman's rank correlation is easy to learn, and doesn't need any assumptions:

  - rank the data in *each* column (not like in the KW test where you rank all columns at once).

  - then use the above formula on the ranks.

  - for a hypothesis test you need tables listing Spearmans's rank critical values.

- See section 19.9, p. 398 of your text

5) *r*, even if significant, DOES NOT IMPLY one variable is "causing" the effect in the other. There is not necessarily any "causation".

a) In Europe, they have found a strong positive correlation (significant) between the number of storks and the number of babies. This is absolutely true!

b) But, obviously (I hope!), storks don't bring babies, so what is going on??

c) Be patient - to be discussed in class.

F) Doing correlations in R:

This is not difficult. Arrange your data in two columns. For the above plant example, we should arrange our data exactly as above (keep in mind that R doesn't like spaces in variable names).

Then we simply do:

```
cor(area,weight)
```

And R gives us:

```
[1] 0.8147143
```

To get the actual correlation test, do:

```
cor.test(area,weight)
```

And the result is:

```
Pearson's product-moment correlation

data:  plant$leafarea and plant$dryweight
t = 4.6599, df = 11, p-value = 0.0006938
alternative hypothesis: true correlation is not equal
                                                    to 0
95 percent confidence interval:
 0.4785463 0.9425797
sample estimates:
     cor
0.8147143
```

And just for completeness sake, here's Spearman's rank correlation:

```
cor.test(area,weight,method = "spearman")

     Spearman's rank correlation rho

data:  plant$leafarea and plant$dryweight
S = 74.6022, p-value = 0.00116
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7950489

Warning message:
In cor.test.default(plant$leafarea, plant$dryweight,
method = "spearman") :
  Cannot compute exact p-values with ties
```