Investigating relationships with Pearson r correlations

Often, when you get a distribution of scores across two variables, you may be interested in determining if they are related to one another in any systematic way. For example, say you have scores on a math test and scores on an English test – are they related? Might it be that the better you do on the English test the better you'll do on the math test? Investigating relationships between two variables is a common analytic procedure and can often give us much good information. Before you get started, review "Measures of relationship" beginning on pg. 425.

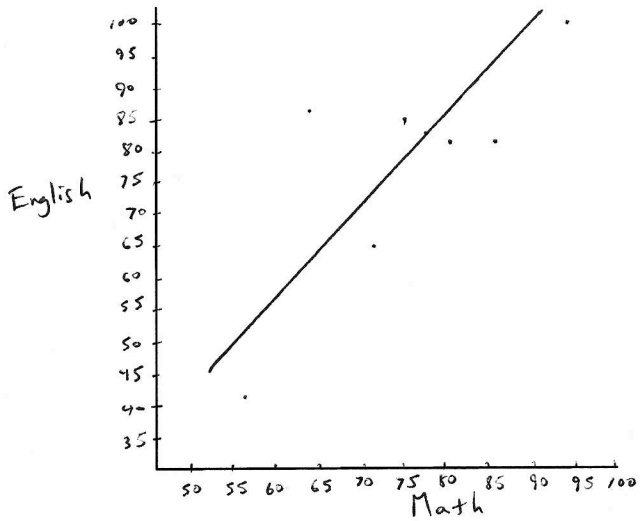
By far, the most common measure of relationship (correlation is a synonym) is the Pearson r – that's the one we're going to talk about here. The Pearson r is restricted to investigating relationships between two continuous variables – meaning, variables that go up or down on a smooth scale rather than at discrete points or intervals. For example, age is a continuous variable while gender is not – it's an "either/or" kind of thing. Likewise, most test scores are continuous variables while hair color is not (usually brown, black, blonde… and so forth) unless you phrase hair color as something like "percentage of brown pigmentation" or something that allows it to smoothly increase.

The other serious caution in correlations is that they do not prove causation. There is, believe it or not, a weak, positive correlation between the height of corn in Nebraska and the murder rate in inner-city Chicago. So, as the height of corn increases in Nebraska, the murder rate also grows. They are related statistically but clearly one does not cause the other – the mediating variables here is, of course, temperature. Hot weather makes corn grow tall and also encourages people to be out on the streets in Chicago… where they bump into each other and then start shooting. Anyway… keep that in mind – correlation says nothing of causation.

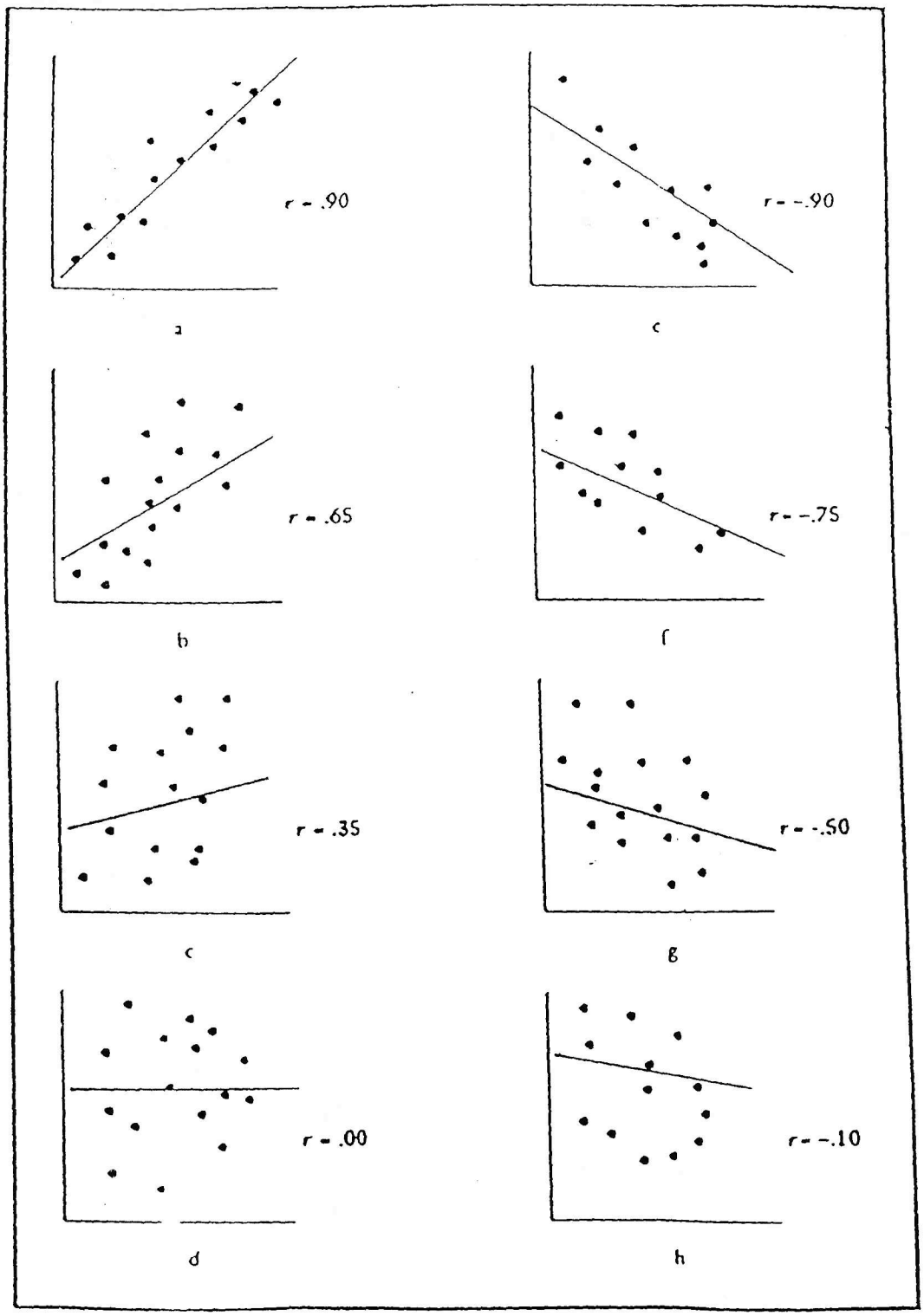
So… here we go on our correlation lesson. One useful and simple thing to do to investigate relationships between two continuous variables is to make a scatterplot and visually investigate their relationship. To make a scatterplot, put one variable along each axis and simply plot the scores. Follow my example then do the height and shoe size one on your own.

Me

	Math scores	English scores
1.	64	86
2.	94	98
3.	42	56
4.	78	82
5.	70	64
6.	82	80
7.	76	84
8.	86	80

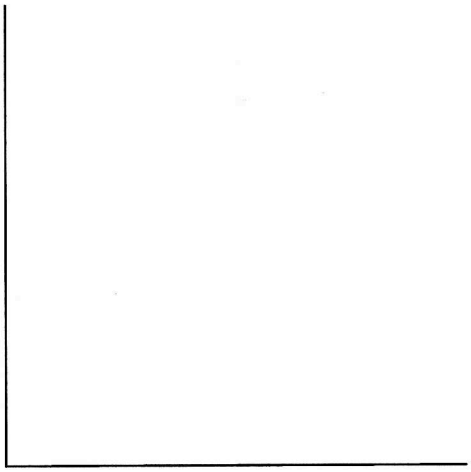


When you're done plotting the data points, the idea is to draw a "best fit" line – a line through the points that leaves about as many above the line as it does below the line. Of course, a perfect positive correlation (+1.00) would have a best fit line slopping at 45 degrees from the lower left hand corner to the upper right hand corner. A perfect negative correlation would have a line sloping at 45 degrees from the upper left hand corner to the lower right – meaning for every one up on the first variable the other variables goes one down – just the opposite of a positive correlation.


r = .90
a
r = -.90
c
r = .65
b
r = -.75
f
r = .35
e
r = -.50
g
r = .00
d
r = -.10
h

You

Height in inches:		Shoe size:
1:	65	6
2:	65	8
3:	67	10
4:	69	10
5:	70	9
6:	71	12
7:	71	11
8:	73	11
9:	74	10
10:	75	13



(Remember, when you plot the numbers they have to both be on the same scale... increasing by 2's or 5's or something)

1. Draw a best fit line and compare it with the scatterplots on the next page. How are your two variables related (positively/negatively, strongly/weekly)? Estimate the strength of the relationship by guessing at a correlation coefficient.

   r =

2. Calculate the correlation coefficient of the height/shoe size data (see pg. 434 -436). Follow my example first:

Me

here's the formula for Pearson r

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

set up a table...

scores	x	y	x^2	y^2	xy
	10	10	100	100	100
	10	8	100	64	80
	8	6	64	36	48
	6	6	36	36	36
	6	4	36	16	24
	$\Sigma x = 40$	$\Sigma y = 34$	$\Sigma x^2 = 272$	$\Sigma y^2 = 252$	$\Sigma xy = 288$

then plug it all in to the formula!

$$r = \frac{288 - \frac{(40)(34)}{5}}{\sqrt{\left[272 - \frac{(40)^2}{5}\right]\left[252 - \frac{(34)^2}{5}\right]}}$$

$r = .51$

whew! :-

Finally, what's cool about correlation coefficients is that we can get significance figures for them. Significance figures give us a level of confidence we might have in our data. You've probably seen them before in research as p <.05 – means the probability (p = probability) of these numbers happening by chance is less than 5 times out of 100. In education, we typically look at probability values of .05 and less, like maybe .01 or .001 – rarely do we see something different from this. So significance value of .001 means there's less than one chance in a thousand that these numbers occurred by random chance or random error – in other words, we have great faith that there's a pretty darn good chance that we should believe the numbers. Significance says nothing about whether or not the relationship is meaningful. That's something you have to decide for yourself – it only tells you the chances that it happened by chance – if you know what I mean! To read more about significance… you're going to have to skip ahead to unit 8 and download the readings on inferential statistics. Read in part 1… pgs. 449-454… the section starting with "tests of significance." Yes, I know these files don't have good markings for page numbers… but start at the first page and number them.

3. When you're ready to continue, download the table of significance for Pearson r (see unit 7 web page) and look your correlation coefficient up on the table. First, find the appropriate degrees of freedom (the formula for df when using Pearson r is df=n-1, where n is the number of pairs. So… your df should be 9… then see if your correlation coefficient is higher than any of the numbers for significance at each of the different probability levels – moving across toward the right). Is your correlation statistically significant? At what level of significance? Write out in English what this really means.

4. Finally, move down on the degrees of freedom column representing an increase in sample size. What happens to the significance figures? Is it easier or harder to get a statistically significant finding with a larger sample size? Why do you think quantitative researchers work so hard to get as big of samples as they can possible get? Pretty cool huh!