XIII. Correlations

Scenario: You will have taken a *pair* of observations on each subject in your sample. Usually each observation is measuring a different (but related) response.

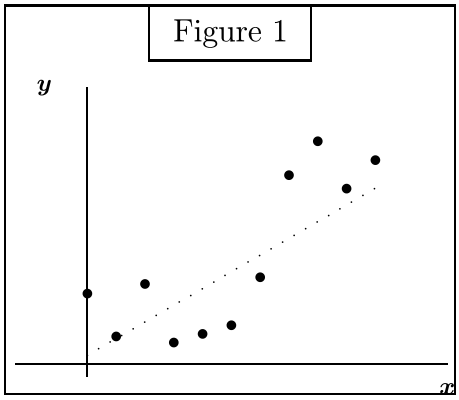
√ The data will look like

first observation	second observation
$\boldsymbol{x}_1$	$\boldsymbol{y}_1$
$\boldsymbol{x}_2$	$\boldsymbol{y}_2$
$\boldsymbol{x}_3$	$\boldsymbol{y}_3$
$\vdots$	$\vdots$
$\boldsymbol{x}_n$	$\boldsymbol{y}_n$

√ *Research Objective.* You will be interested in seeing if the data exhibit a straight line relationship. In Figure 1 the data all fall approximately on the dotted line.

√ *Graphical Presentation.* Sometimes it will be useful to present the data graphically in a *scatter plot*. You should plot the $(\boldsymbol{x}, \boldsymbol{y})$ pairs to get a graph which will look something like:

Figure 1



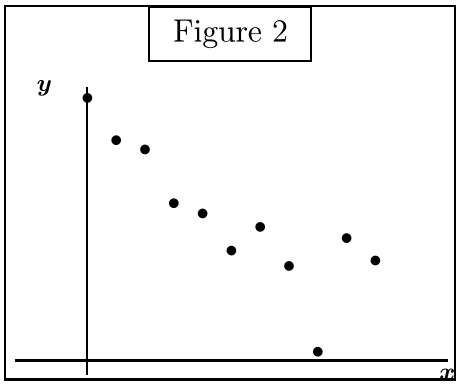
The graphical presentation can help you decide if the responses measured by $\boldsymbol{x}$ and

y are correlated (related) according to a straight line formula:

$$y = mx + b.$$

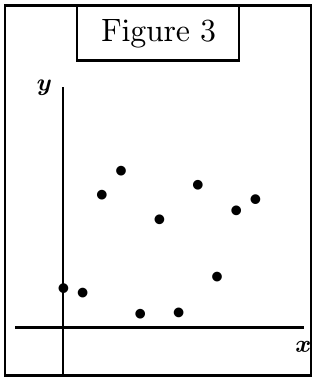
The parameter m will measure the *slope* (or slant) of the line; the parameter b measures where the line intercepts the y axis. Eventually, we will use the correlation coefficient to find the parameters m and b. Figure 1 above shows a *positive* correlation since increases in x seem to imply increases in y (corresponding to a positive value for m). Figure 2 below shows a *negative* correlation since increases in x seem to imply decreases in y (corresponding to a negative value for m).

Figure 2



Sometimes the scatter plot will suggest that there is no relationship between the x and y observations. Figure three shows a scatter plot of uncorrelated data.

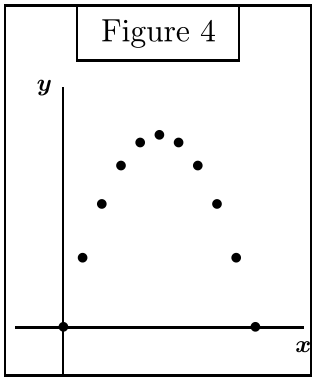
Figure 3



Other times the relationship revealed by the scatter plot might be nonlinear. In this

case the data are also uncorrelated. Figure 4 represents such a situation.

Figure 4



✓ *Numerical Calculations.* The *correlation coefficient* r measures how well the data fit a straight line plot. The correlation coefficient has the following properties:

- The correlation coefficient is always between -1 and +1.
- If the correlation coefficient is zero the data are uncorrelated.
- If the correlation coefficient is +1 the data are perfectly positively correlated.
- If the correlation coefficient is -1 the data are perfectly negatively correlated.

In general, you will never get the ideal situations $r = 0$, $r = 1$ or $r = -1$. Instead you will get mixed results (like $r = 0.473$) which suggest some relationship between the variables but not a perfect one. In this section we will concentrate on how to compute the correlation coefficient and give a "rule-of-thumb" interpretation for the result. In a later section we will discuss hypothesis testing for correlation coefficients.

How to find the correlation coefficient.

Solution Template

Step 1. First make a column in the data table for the products of the x and y

observations.

first observation	second observation	products
x_1	y_1	$x_y \times y_1$
x_2	y_2	$x_y \times y_1$
x_3	y_3	$x_y \times y_1$
$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$x_y \times y_1$

Step 2. Find the means for all three columns and find the *population* standard deviations for the first two columns. *Use the population standard deviation even if you have sample data!* You will be finding:

	1st obs	2nd obs	products
means	$\bar{x}$	$\bar{y}$	average of products
standard deviations	s_x	s_y	–

Step 3. The correlation coefficient is then given by the formula

$$r = \frac{\text{average of products} - \bar{x}\bar{y}}{s_x s_y}$$

End of Solution Template

Interpretation. Some of the variability in the y's is due to the dependence on the x's and some is due to other factors (such as measurement error). The square of the correlation coefficient (r^2) tells you approximately what proportion of the variability in the y's is due to the dependence on the x's. (The precise formula for the variation in y due to x is

$$s_y^2\left(\frac{n}{n-2}r^2 - \frac{2}{n-2}\right)$$

For large values of n this is approximately $s_y^2 r^2$. The proportion of variability in y to the effect measured by x is

$$\frac{\text{variability in } y \text{ due to } x}{\text{total variability in } y} = \frac{s_y^2\left(\frac{n}{n-2}r^2 - \frac{2}{n-2}\right)}{s_y^2}$$
$$= \frac{n}{n-2}r^2 - \frac{2}{n-2}$$

which, for large values of n, is approximately r^2.)

Example. A counselor is treating individuals with an eating disorder by administering a drug which enhances appetite. The researcher administers 11 subjects different dosages of the drug for 30 days and then measures the net change in weight. The drug dosages (measured in mg) and the change in weight (measure in pounds) are listed below.

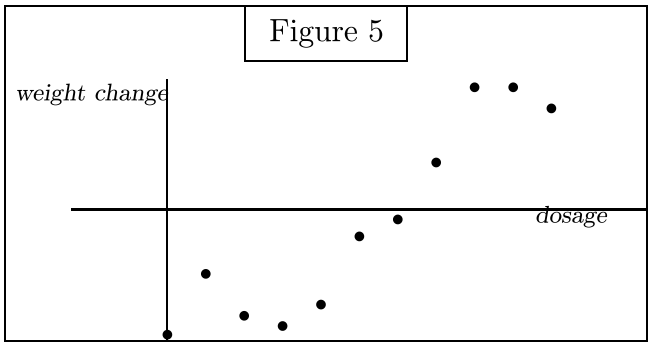
dose	weight change
0.000	-4.368
0.100	-2.255
0.200	-3.714
0.300	-4.063
0.400	-3.324
0.500	-0.960
0.600	-0.369
0.700	1.615
0.800	4.216
0.900	4.224
1.000	3.492

Find the correlation coefficient between dosage and weight change.

Solution. .

To help visualize the data, we will first draw a scatter plot:

Figure 5



The data suggest that there is some positive relationship between dosage and weight loss, but it is not clear how strong the relationship is. To help measure this, we will compute the correlation coefficient.

Step 1. Augment the data table with a column for products.

dose	weight change	products
0.000	-4.368	
0.100	-2.255	
0.200	-3.714	
0.300	-4.063	
0.400	-3.324	
0.500	-0.960	
0.600	-0.369	
0.700	1.615	
0.800	4.216	
0.900	4.224	
1.000	3.492	

Next fill in the third column with the product of the first two.

dose	weight change	products
0.000	-4.368	0.000
0.100	-2.255	-0.225
0.200	-3.714	-0.743
0.300	-4.063	-1.219
0.400	-3.324	-1.330
0.500	-0.960	-0.480
0.600	-0.369	-0.221
0.700	1.615	1.130
0.800	4.216	3.373
0.900	4.224	3.802
1.000	3.492	3.492

Step 2. Find the averages for all three columns and the standard deviations for the first two columns.

	dose	gain/loss	products
means	$\bar{x} = 0.500$	$\bar{y} = -0.500$	0.689
stand. dev.	$s_x = 0.316$	$s_y = 3.223$	

Step 3. Find the correlation coefficient.

$$\begin{aligned} r &= \frac{\text{average of products} - \bar{x}\bar{y}}{s_x s_y} \\ &= \frac{0.689 - (0.500)(-0.500)}{(0.316)(3.223)} \\ &= \frac{0.939}{1.0185} \\ &= 0.922 \end{aligned}$$

Thus the correlation coefficient is $r = 0.922$. The square of the correlation coefficient $r^2 = 0.850$ and so about 85% of the variability in weight loss can be attributed to the differences in dosage. ∎

Question. If only 85% of the variability in weight loss can be attributed to dosage, what other factors might contribute to weight loss?

Problems

1. Find the correlation between the following sets of x and y values.

x	y
4	1
5	2
3	4
6	6
7	7

2. A researcher administers a writing test to men and women and obtains the following scores:

Men	Women
93	94
90	70
88	69
66	68
65	60
61	41
60	

Is there a correlation between gender and the score on this test? (Hint: assign a number to each gender, for example, assign a "0" to "male" and a "1" to "female" before attempting to find the correlation.)