Chapter 12
Correlation

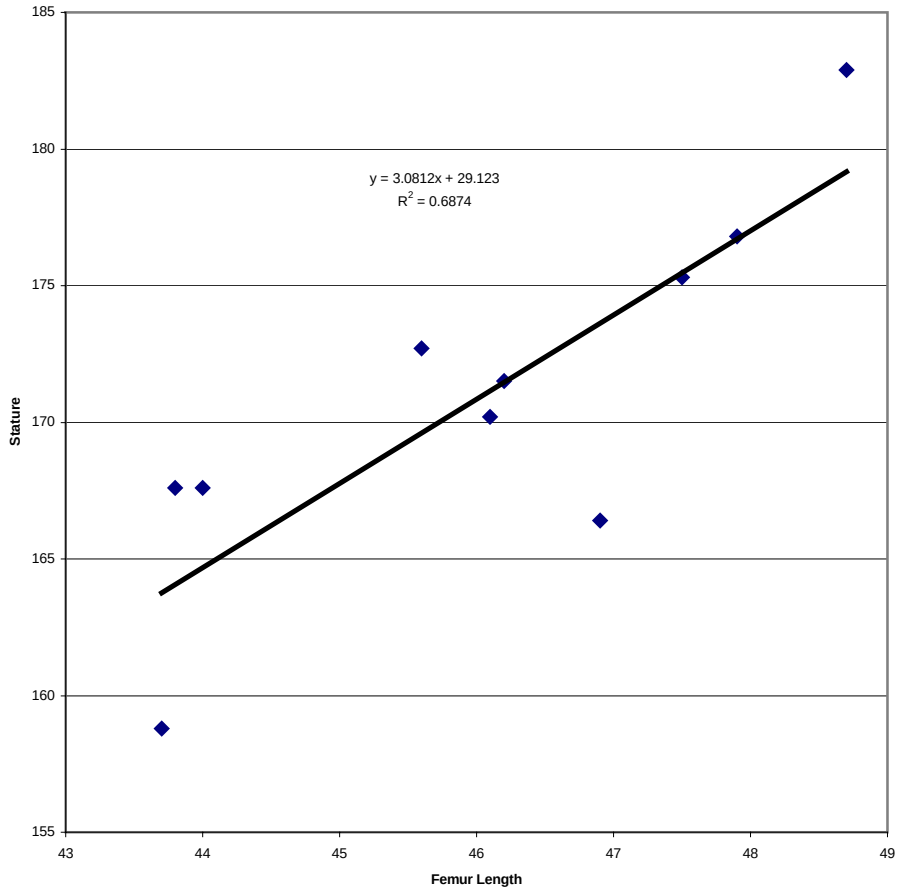
Correlation is very similar to regression with one very important difference. Regression is used to explore the relationship between an *independent* variable and a *dependent* variable, whereas correlation is used to consider a relationship between two dependent variables. To illustrate this difference, let us consider the relationships between stature, femur lengths, and humerus lengths on ten men from the University of New Mexico documented skeletal collection. These data are presented in Table 12.1.

Table 12.1. Stature and long bone lengths on 10 males from the UNM documented skeletal collection.

Stature	Femur length	Humerus Length
176.8	47.9	34.6
172.7	45.6	31.4
158.8	43.7	31.4
166.4	46.9	34
182.9	48.7	35
167.6	43.8	31.3
170.2	46.1	33.9
167.6	44	32.3
171.5	46.2	33.6
175.3	47.5	33.5

We know that stature is, at least in part, a function of the length of the femur. We can therefore explore the relationship between the independent X variable, femur length, and the dependent variable Y, stature. Figure 12.1 presents this relationship graphically, as well as the regression equation and r^2.

Figure 12.1. Stature and long bone lengths of ten males.

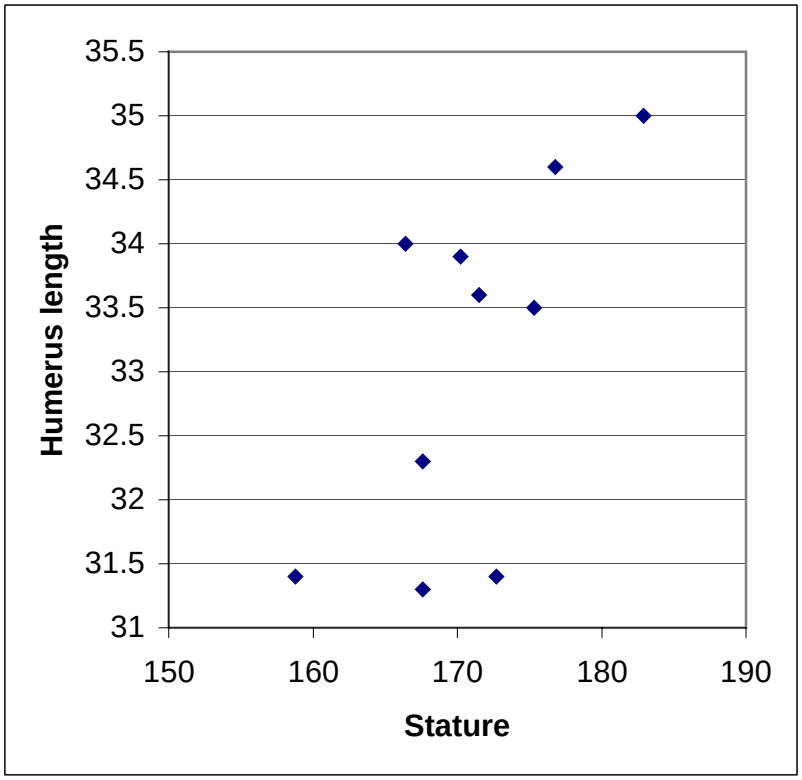


Here we see that there is indeed a linear relationship between femur length and stature, and that we could predict a new Y for any given X.

What about the relationship between stature and humerus length? Is stature a function of humerus length? No, the length of one's upper arm contributes in no way to overall stature. These two variables, however, may be correlated, both being a function of one or more additional variables. Stature and femur lengths may both be the products of factors such as genetics and nutrition for example, as well as a wide array of other environmental factors. We therefore might expect a connection between stature and humerous length such that individuals with large statures would also tend to have large humerus lengths. That is, the two variables might be correlated with each other because they both are the

products of underlying shared variable(s). Correlation is the appropriate statistical procedure with which to explore this possibility. Figure 12.2. presents the relationship visually.

Figure 12.2. Relationship between humerus length and stature.



We see that there is a general positive linear relationship, but how do we estimate the strength of it? This is determined by computing *r*, the Pearson's product moment correlation coefficient. To begin calculation, note that we no longer have an independent variable X and a *dependent* variable Y, but rather two *dependent* variables Y_1 and Y_2. Therefore, it doesn't matter which variable is depicted on the horizontal or vertical axes.

To calculate correlation, we solve for the following quantities, where Y_1 = stature and Y_2 = humerus length.

Quantity 1: $\sum Y_1 = 1709.8$

Quantity 2: $\sum Y_1^2$ =292732.2

Quantity 3: $\sum Y_2$ =331

Quantity 4: $\sum Y_2^2$ =10973.48

Quantity 5: $\sum Y_1 Y_2$ =56649.57

Quantity 6, the Sum of Squares of Y_1:

$$\sum y_1^2 = \sum Y_1^2 - \frac{(\sum Y_1)^2}{n} = \text{Quantity 2} - \frac{(\text{Quantity 1})^2}{n} = 292732.3 - \frac{(1709.8)^2}{10} = 390.596$$

Quantity 7, the Sum of Squares of Y_2:

$$\sum y_2^2 = \sum Y_2^2 - \frac{(\sum Y_1)^2}{n} = \text{Quantity 4} - \frac{(\text{Quantity 3})^2}{n} = 10973.48 - \frac{(311)^2}{10} = 17.38$$

Quantity 8, the Sum of Products:

$$\sum y_1 y_2 = \sum Y_1 Y_2 - \frac{(\sum Y_1)(\sum Y_2)}{n} = \text{Quantity 5} - \frac{(\text{Quantity 1})(\text{Quantity 3})}{n} = 56649.57 - \frac{(1709.8)(331)}{10} - 55.$$

Quantity 9, the Pearson's Product Moment Correlation Coefficient:

$$r_{1,2} = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2 \sum y_2^2}} = \frac{\text{Quantity 8}}{\sqrt{(\text{Quantity 6})(\text{Quantity 7})}} = \frac{55.19}{\sqrt{(390.596)(17.38)}} = .6698$$

Pearson's *r* can range from –1, a perfect negative correlation in which one variable decreases as the other increases, to 1, a perfect positive correlation in which both variables increase. Values close to zero indicate that there is no correlation between the two variables of interest. Yet, how close to zero is close enough to indicate no relationship?

A significance test tells us the answer. In correlation, we are actually testing the null hypothesis: H_o:ρ=0, where ρ (vocalized as rho or roe) is the population parameter of the correlation coefficient, *r*. We use a t-test to determine if our *r* value is significantly different from 0 in standard deviation units to determine whether or not there is a significant correlation. We set the level of rejection (alpha) at .05.

The standard error of the correlation coefficient is:

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

The t test:

$$t = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = r \times \sqrt{\frac{n-2}{1-r^2}} = .6698\sqrt{\frac{10-2}{1-.4486}} = 2.5513$$

As $t_{.05[8]} = 2.306$, we reject H_0. The correlation is significant.

Spearman's Rank Order Correlation Coefficient

Many times we wish to examine relationships between variables that are not ratio or interval scales of measurement. Instead, we may have ordinal level data that allows us to be certain only of the order among variates, not distance. In these situations we can use Spearman's Rank Order Correlation Coefficient to explore the relationship between ordinal variables.

Consider the following data presented in Table 12.2. These are rank order fish bone abundances from two sites in the lower Illinois Valley presented by Bonnie Styles (1995).

Taxon	Newbridge %	Newbridge rank (R_1)	Carlin %	Carlin rank (R_2)	(R_1-R_2)	(R_1-R_2)^2
Bullhead	33.96	1	11.27	3	-2	4
Bowfin	21.05	2	16.67	2	0	0
Buffalo	12.90	3	10.29	4.5	-1.5	2.25
River Catfish	10.53	4	19.12	1	3	9
Bass	6.11	5	10.29	4.5	.5	.25
Sunfish	4.41	6	0.98	9	-3	9
Pike	3.40	7	3.92	8	-1	1
Redhorse	3.06	8	5.88	7	1	1
Freshwater drum	2.38	9	9.80	6	3	9
Crappie	.51	10	0.49	10	0	0
Σ						35.50

While fish bone counts are continuous, different sizes of the fish and differences in bone preservation, among other factors, require us to be suspicious of the actual counts. We really aren't sure if the actual counts are truly representative of the population of interest. However, we might be interested in knowing whether there is a relationship, either positive or negative, between the frequencies of various fish species such that sites with a the remains of one type of fish also are likely (or unlikely) to have the remains of another type of fish. Further, we might be comfortable with the relative frequencies of the fish bones as a reflection of their use at the sites, even if we don't think these frequencies accurately reflect the total number of fish used at the site. As a result, we can accurately create an ordinal measure of the relative percentages of the fish in each site.

To determine if there is a correlation between the relative frequencies of fish at the Newbridge and Carlin sites, we can use the Spearman's rank order correlation coefficient rather than Pearson's. It is calculated as follows:

Spearman's *r* is calculated below:

$$r_s = 1 - \frac{6\sum (R_1 - R_2)^2}{n(n^2 - 1)} = 1 - \frac{6(35.5)}{10(100 - 1)} = 1 - \frac{213}{990} = .7848$$

Values of Spearman's *r* are interpreted in the same manner as Pearson's. Values of –1 indicate a perfect negative correlation, values of 1 indicate a perfect positive correlation. Values close to zero indicate no relationship. The formula $t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}}$ for the t-test introduced above is used to evaluate the significance of Spearman's *r* when n>10. Critical values for n≤10 are presented in Table 12.3.

Table 12.3. Critical values for Spearman's *r* when n≤10.

N	Significance level (one-tailed test)	
	.05	.01
4	1.000	
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746

The null hypothesis is $H_0 : r = 0$ and the level of rejection (alpha) is set at .05. Comparing our Spearman's *r* to the critical value listed in Table 12.3, we find that we must reject the null hypothesis and can conclude that there is a positive correlation between the fish assemblages from the Newbridge and Carlin sites.

Ethnobotanical analysis also lends itself to Spearman's *r*. Styles (1985) presents data on botanical samples from the Newbridge and Carlin sites. These data are presented in table 12.4. We can evaluate the null hypothesis $H_0 : r = 0$ to determine if there is a correlation in the relative frequencies of the various botanical remains between the sites. The level of rejection (alpha) is set at .05.

Table 12.4. Ethnobotanical data for the Newbridge and Carlin sites.

Seed Group	Newbridge %	Newbridge rank (R_1)	Carlin %	Carlin rank (R_2)	(R_1-R_2)	(R_1-R_2)^2
Starchy cultivated (?)	84.82	1	89.31	1	0	0
Misc.	10.75	2	4.47	3	-1	1
Oily cultivated	2.80	3	.45	5	-2	4
Starchy non-cultivated	1.41	4	4.52	2	2	4
Sumac	.09	5	.73	4	1	1
Fleshy fruits	.08	6	.23	7	-1	1
Weed seeds	.07	7	.24	6	1	1
N of seeds	15,009		2,868			
Σ						12

We calculate Spearman's r as follows:

$$r_s = 1 - \frac{6\sum(R_1 - R_2)^2}{n(n^2 - 1)} = 1 - \frac{6(12)}{7(49 - 1)} = 1 - \frac{72}{336} = .7857$$

We see that there is a positive relationship, and that the structure of the botanical assemblages of the sites of Newbridge and Carlin are similar. Comparing r_s to the critical value of .714 listed in Table 12.3 for N=8 indicates that this correlation is significant.

Now let us compare Newbridge and another site, Weitzer. These date are presented in Table 12.5. Again we testing the hypothesis $H_0 : r = 0$ and the level of rejection (alpha) is set at .05.

Table 12.5. Paleobotanical information from the Newbridge and Weitzer sites.

Seed Group	Newbridge %	Newbridge rank (R_1)	Weitzer %	Weitzer rank (R_2)	(R_1-R_2)	(R_1-R_2)^2
Starchy cultivated (?)	84.82	1	26.67	2	-1	1
Misc.	10.75	2	48.33	1	1	1
Oily cultivated	2.80	3	0	6.5	-3.5	12.25
Starchy non-cultivated	1.41	4	6.67	4.5	-.5	.25
Sumac	.09	5	11.67	3	2	4
Fleshy fruits	.08	6	6.67	4.5	1.5	2.25
Weed seeds	.07	7	0	6.5	.5	.25
N of seeds	15,009		2,868			
Σ						21

Spearman's r is calculated below:

$$r_s = 1 - \frac{6\sum (R_1 - R_2)^2}{n(n^2 - 1)} = 1 - \frac{6(21)}{7(49 - 1)} = 1 - \frac{126}{336} = .375$$

We see that this relationship is much weaker that that between Newbridge and Carlin. Comparing r_s to the critical value of .714 listed in Table 12.3 for N=8 indicates that this correlation is not significant, and that we cannot reject the null hypothesis.

Pearson's Product Moment Correlation Coefficient and Spearman's Rank Order Correlation Coefficient provide excellent tools for examining relationships between variables that are not linked in any causal ways. Similarly, we now proceed with an examination of association between two or more categorical variables. That is the subject of Chapter 13.