22

Correlation Coefficients

The Meaning of Correlation
Correlation and Data Types
Pearson's r
Spearman rho
Other Coefficients of Note
Coefficient of Determination r^2

The concept of correlation was introduced in Chapters 1 and 5. Our focus since Chapter 16 has been basic statistical procedures that measure ***differences between groups*** -- one-sample, two-sample, and *k*-sample tests.

Now we turn our attention to basic statistical procedures that measure the ***degree of association between variables***.

Dr. Wesley Black studied the relationship between rankings of selected learning objectives in a youth discipleship taxonomy between full-time church staff youth ministers and seminary students enrolled in youth education courses at Southwestern Seminary.[1] Questionnaires were returned by 318 students and 184 youth ministers.[2] Ten objectives in each of five categories (Personal Ministry, Christian Theology and Baptist Doctrine, Christian Ethics, Baptist Heritage, and Church Polity and Organization) were ranked by these two groups.

The basic question raised by Black in this study was whether students prioritized discipleship training objectives for youth in the same way as full-time ministers in the field. Using the Spearman rho correlation coefficient, Black found the correlations of rankings generated by students and ministers of the ten items for each of five categories were as follows: Personal Ministry, 0.915; Christian Theology and Baptist Doctrine, 0.867; Christian Ethics, 0.939; Baptist Heritage, 0.939; and Church Polity and Organization, 0.927.

Each of these are strong positive correlations[3] between the rankings of objectives by students and ministers.

---

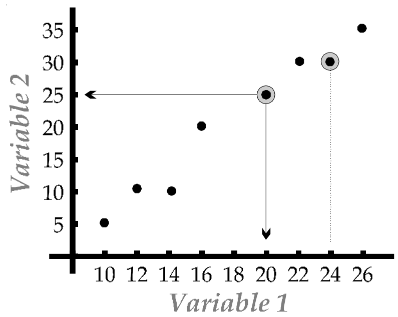
[1]Wesley Black, "A Comparison of Responses to Learning Objectives for Youth Discipleship Training from Ministers of Youth in Southern Baptist Churches and Students Enrolled in Youth Education Courses at Southwestern Baptist Theological Seminary," (Fort Worth, Texas: Southwestern Baptist Theological Seminary, 1985).

[2]Black received 356 responses from students, but 38 of these were full-time youth ministers, and so were excluded from the study, leaving 318 student responses. Of the 307 responses from youth ministers, 197 indicated they were "full-time church staff youth ministers." Thirteen additional responses were eliminated for incompleteness, leaving 184 youth minister responses. pp. 71-72

[3]Any coefficient over 0.80 indicates a strong positive correlation.

The Meaning of Correlation

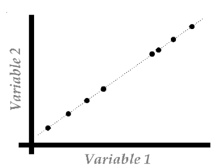
When we discussed the frequency distribution (chapter 15), we plotted *values* of X on the x-axis, and the *frequency* of the X-values on the y-axis. In this plot, there was one score per subject.



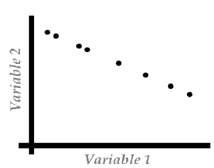
In graphing a correlation between two variables, there are ***two scores* per subject** — an X-score and a Y-score. We plot the X-scores on the x-axis and Y-scores on the y-axis. A single dot represents the intersection between each X-Y pair. Notice the diagram to the left. The single point in the shaded circle represents two scores from a single subject in a study: a ***20*** on variable 1 and a ***25*** on variable 2.

Notice how the dots form a pattern in two-dimensional space. The tighter the pattern, the higher the correlation. The looser the pattern, the lower the correlation.

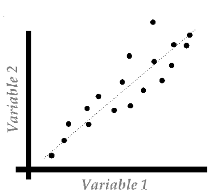
These patterns are called ***scatterplots***. The scatterplots at left illustrate various kinds of associations. The first scatterplot shows a *perfect positive correlation.* The correlation is positive because **variable 2 increases as variable 1 increases**. It is a *perfect* correlation because all the points fall on a straight line. (The line has been included in this diagram, but is not part of the scatterplot.)



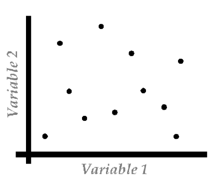
The second scatterplot shows a *perfect negative correlation*. The correlation is negative because **variable 2 *decreases* as variable 1 increases**. It is *perfect* because all the points fall on a straight line (not shown in this diagram).



The third scatterplot shows a *moderately positive correlation*. Notice how most of the data points do not fall on the line. It is a moderate correlation, however, because the points fall in a tight pattern around the line. Notice the pattern is ***linear*** -- that is, a pattern suggesting a line.



The fourth scatterplot shows *no correlation.* Scores on one variable have no systematic association with scores on the other. The scatterplot presents ***no linear pattern*** among the points.



Beyond the graphical representation of association, we can mathematically compute the *degree of association* between two variables. The numerical result of such a computation is called a ***correlation coefficient***. The value of these coefficients usually range from -1.00 to +1.00.

A ***positive*** coefficient indicates that two variables systematically vary in the same direction: ***as one variable increases, the other variable tends to increase.*** The closer the coefficient is to +1.00, the stronger the positive association.

A ***negative*** coefficient indicates that two variables systematically vary in opposite directions: ***as one variable increases, the other variable tends to decrease.*** The closer the coefficient is to -1.00, the stronger the negative association.

A coefficient ***close to zero*** indicates that no systematic co-varying exists between the variables. There are several important correlation procedures. They differ according to the data types of the variables.

Correlation and Data Types

Since chapter 16, we have focused on interval or ratio data types. In this chapter, we broaden our focus. There are correlational procedures for all four data types (nominal, ordinal, interval, ratio).

The ***Pearson's Product Moment Correlation Coefficient*** (r_{xy}) computes the correlation between two interval or ratio variables. ***Spearman's rho*** (r_s) computes the correlation between two ordinal, or ranked, variables. The ***Contingency Coefficient (C)*** and ***Cramer's Phi*** (ϕ_C) compute the strength of relationship when testing nominal data analyzed by a χI procedure (Chapter 23). The ***Phi Coefficient*** (r_ϕ) computes the correlation between two ***dichotomous variables*** (two and only two categories (Yes or No, True or False).

Additionally, a study may require the computation of a correlation coefficient between mixed data types. ***Point biserial*** is used when one variable is interval/ratio and the second is dichotomous. ***Rank biserial*** is used when one variable is ordinal and the second is dichotomous. We can summarize these various coefficients like this:

	Variable 1			
Variable 2	**Interval/Ratio**	**Ordinal**	**Nominal**	**Dichotomous**
Interval/Ratio	r_{xy}	r_s*		**Point Biserial**
Ordinal	r_s*	r_s		**Rank Biserial**
Nominal			C, ϕ_C**	
Dichotomous	**Point Biserial**	**Rank Biserial**		r_ϕ

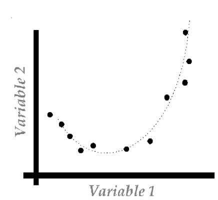
**requires interval/ratio data to be ranked*

***Requires χ^2 value*

Finally, *Kendall's Coefficient of Concordance (W)* computes the correlation of three or more rankings of of items. Now we'll look at how each of these correlation coefficients are computed.

Pearson's Product Moment Correlation Coefficient (r_{xy})

The most popular correlation coefficient is the Product Moment correlation coefficient, better known as ***Pearson's r***. Pearson's r is used to determine the correlation between two variables under three conditions.



First, both variables must be **interval or ratio measures** (i.e. attitude scales, test scores).

Second, the relationship between the two variables must be **linear** — the data points must generally fall along a straight line. A non-linear relationship between variables, shown at right, produces a Pearson's r near zero, even though is it clear from the example that there is a strong ("quadratic," of the type $y=x^2$) relationship between the two variables.

The third condition is that both variables are **normally distributed**. A skewed distribution produces a smaller r than a normal distribution.

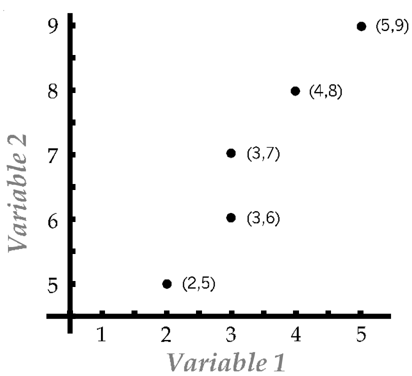
Use a large scale for variables in correlational analysis, since the larger the variability, the stronger the coefficient will be. **A common mistake** in research design is to use age categories rather than actual ages, or salary categories rather than actual dollar values. The range of categories will always be much smaller than the range of actual data, reducing the value of r.

Pearson's r is computed with the following formula:

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n(\Sigma X^2)-(\Sigma X)^2][n(\Sigma Y^2)-(\Sigma Y)^2]}}$$

where n equals the number of score-pairs, and X and Y equal paired scores. While this formula is somewhat foreboding, it consists of the following simple components:

ΣXY	Multiply X by Y and sum
ΣX	Sum all the X scores
ΣY	Sum all the Y scores
ΣX^2	Square all X's and sum
ΣY^2	Square all Y's and sum
$(\Sigma X)^2$	Square the sum of X
$(\Sigma Y)^2$	Square the sum of Y



Let's say we have a set of 5 paired scores: (3,6), (5,9), (2,5), (3,7), and (4,8). A scatter-plot of this data is shown at left. From what you can see in this graph, do you predict a strong or weak correlation coefficient?

We've put the paired X-Y values in the chart below to facilitate computing the various elements of the Pearson's r formula. The letters in the chart (A-G) refer to the step below (A-G).

X^2	**X**	XY	**Y**	Y^2
====	===	====	===	====
9	**3**	18	**6**	36
25	**5**	45	**9**	81
4	**2**	10	**5**	25
9	**3**	21	**7**	49
16	**4**	32	**8**	64
—	—	—	—-	—
63	17	126	35	255
ΣX^2	ΣX	ΣXY	ΣY	ΣY^2
F	A	E	C	G
	289		1225	
	$(\Sigma X)^2$		$(\Sigma Y)^2$	
	B		D	

A. Add up the Xs. This is ΣX, and equals 17

B. $(\Sigma X)^2 = 17 \times 17 = 289$
C. Add up the Y's. This is ΣY, and equals 35
D. $(\Sigma Y)^2 = 35 \times 35 = 1225$
E. Multiply each XY pair together and add. This is (ΣXY), and equals 126
F. Square each X and add up the squared values. $\Sigma X^2 = 63$
G. Square each Y and add up the squared values. $\Sigma Y^2 = 255$

Now substituting into the raw score equation we have: *Before going on, be sure to identify each term in the equation above with the chart above and the equation on the previous page.*

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n(\Sigma X^2) - (\Sigma X)^2][n(\Sigma Y^2) - (\Sigma Y)^2]}} = \frac{5(126) - (17)(35)}{\sqrt{[5(63) - (289)][5(255) - (1225)]}}$$

$$= \frac{630 - 595}{\sqrt{[315 - 289][1275 - 1225]}} = \frac{35}{\sqrt{[26][50]}} = \frac{35}{36.056} = 0.971$$

The Pearson r value of +0.971 indicates a very strong -- nearly perfect -- positive correlation between these two variables.

Spearman's rho Correlation Coefficient (r_s)

Spearman's rho yields a correlation coefficient between two ordinal, or ranked, variables.[4] The formula is:

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$

where D is the difference between paired ranks. The number "6" is a constant.

Suppose a pastor asked two staff members to rank ten church objectives according to how well they were being accomplished by the church. Here are the rankings of the ministers.

Objective	*Min Ed Rank*	*Min Youth Rank*
1	1	2
2	2	1
3	3	5
4	4	3
5	5	7
6	6	6
7	7	4
8	8	10
9	9	9
10	10	8

Question: Do these two staff members agree in their evaluation of the objectives?

[4]The "two ranked variables" in Dr. Black's dissertation (p. 1) was "ranking of objectives by students" and "ranking of objectives by youth ministers." This was accomplished by assigning a score value to each individual subject's set of ranks, computing means of these scores, and rank ordering the objectives by the means. The result was a separate ranking high to low of objectives by the two groups. Spearman rho was then applied to compute the degree of agreement between the two rankings.

What is the strength of their agreement?

First we compute the differences (D) between ranks, then square the differences (D^2), sum the squares (D^2), and substitute into the formula. The table below summarizes the process:

Objective	Min Ed Rank		Min Youth Rank		D	D^2
1	1	-	2	=	-1	1
2	2		1		+1	1
3	3		5		-2	4
4	4		3		+1	1
5	5		7		-2	4
6	6		6		0	0
7	7		4		+3	9
8	8		10		-2	4
9	9		9		0	0
n= **10**	10		8		+2	4
					— —	— —
					$\Sigma D=0$	$\Sigma D^2=28$

Objective 1 was ranked highest (1) by the minister of education and second (2) by the minister of youth. Subtracting 2 from 1 yields a difference (D) of **-1**. Squaring D yields a D^2 of **1**. Notice that the sum of differences (D) equals 0.

Summing the D^2 values, we get **28**.

Substituting the value of $\mathbf{D^2}$ and **n** into the Spearman formula, we have

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2-1)} = 1 - \frac{6(28)}{10(99)} = 1 - 0.17 = 0.83$$

The coefficient of +0.83 indicates a strong agreement between the two staff ministers with respect to the rankings of church objectives.

Other Important Correlation Coefficients

Several other correlation coefficients will be mentioned. These will be described but not illustrated by example in the interest of space.

Point Biserial Coefficient

The point biserial correlation coefficient is computed between one interval or ratio variable and one dichotomous variable. The term "biserial" refers to the fact that there are two groups of persons (X= 0,1) being observed on the continuous variable (Y).

Use this procedure to test correlations of attitude scores or test scores of subjects between "haves" and "have-nots": ministers who graduated from seminary and those who did not, preschoolers who have had a specified early education program and those who haven't, staff members who have a specified evaluation procedure and those who do not, and so forth.

Rank Biserial Coefficient

The rank biserial correlation coefficient is much like the point biserial just discussed, except that it uses an ***ordinal variable*** in place of an interval/ratio variable. The coefficient measures degree of relationship between a dichotomous condition (1,0) and a ranking.

Phi Coefficient (r_ϕ)

The **Phi Coefficient** measures the strength of relationship between two dichotomous variables. A study of marital status and attrition rate in college might arbitrarily assign a "1" to married and "0" to not married; a "1" to dropped out and a "0" to remaining in school. Any type of variable that can be classified "1" and "0" can use the phi coefficient.

A positive correlation indicates those who score "1" on one variable tend to score "1" on the other. Using the example above, a positive correlation would mean that married students (1) tend to drop out of school (1) more than unmarried students.

Kendall's Coefficient of Concordance (W)

Kendall's W extends Spearman Rho to more than two groups. This procedure is useful for studies in which three or more groups create rankings of items. The resulting statistic represents the level of agreement among the groups in ranking the items.

For example, you could create a list of ten competencies required of an effective minister of education. You then send this list to a sample of pastors, a sample of ministers of education, and a sample of seminary professors. You ask them to rank order the list. Convert the individual rankings to score values and compute means for each item for each of the three groups. Use the means to rank order the items for each group. Then use Kendall's W to measure the degree of agreement exists among these three groups concerning the relative importance of the stated competencies for ministers of education.

The Coefficient of Determination (r^2)

I've said nothing about significance tests for correlation coefficients. That is to say, I've not suggested a statistic to determine whether a coefficient is "significant" or not, such as: *It is the hypothesis of this study that there is a **significant** positive correlation between number of study hours and grade point average in first year seminary students.*

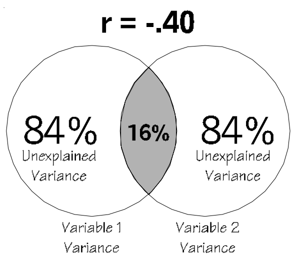
You will find textbooks that provide this information, but it is best left alone. **Leland Wilkerson**, developer of ***SYSTAT: A System For Statistics***, warns against the "smoke and mirrors" of significance testing for correlation coefficients. The null condition for correlation coefficients is H_0: r=0. A "significant correlation" is one that differs from "0." The size of r is directly related to n, the number of subjects. The larger the group, the more likely one will achieve a "significant r," **even if the correlation is meaningless.**

I once helped a Catholic nun analyze correlational data on attitudes of Catholic parents toward sex education in parochial schools. Most of her correlations were very small -- 0.05 to 0.15 -- yet her SPSS printouts declared all of them "significant."

The reason for this was the number of subjects in her study -- nearly 500 couples, almost 1000 subjects. ***What does a "significant" correlation of r=0.07 mean? Very little.***

Here we again see the distinction between statistical significance and practical importance.

A different approach, a more meaningful approach, in determining the importance of a correlation coefficient, is the ***coefficient of determination (rI)***. By squaring the correlation coefficient, one obtains a measure of the *common variance* between two variables, the **proportion of variance accounted for** in one of the variables, or "explained" by, the other. If the correlation between "marital satisfaction" and "number of months married" is -0.40, then 16% of the variance (-.40 x -.40 = .16) of one variable is "accounted for" by the variance of the other (the shaded area at right). We could say that 16% of the variability in marital satisfaction and number-of-months-married "overlaps". It follows that 84% of the variability is unaccounted for.



In the Catholic sex education study mentioned above, the r^2 value of r=0.07 is 0.0049, or 0.49%: one-half of one percent of variance accounted for. Ninety-nine and three-fourths percent (99.51%) of variance was unaccounted for. This was a meaningless "significant" finding to be sure.

We will use the concept of r^2 much more when we discuss regression analysis.

Summary

In this chapter you have been introduced to the concept of correlation. You have learned how to compute the two most popular correlation coefficients, Pearson's r and Spearman rho, as well as learned of several other helpful correlational tools. You have been introduced to the coefficient of determination (r^2) which will be of central importance in Chapter 26, Regression Analysis.

Vocabulary

Coefficient of Determination	**proportion** of variance in one variable **accounted for** by another (r^2)
Contingency Coefficient	measure of **association between two nominal variables** (max < 1)
Correlation	degree of **association between two variables**
Correlation coefficient	numerical measure of degree of **association between two variables**
Cramer's Phi	measure of **association between two nominal variables** (max = 1)
Kendall's tau	measure of association between **two sets of ranks** (n < 10 pairs)
Kendall's W	measure of association among **three+ sets of ranks**
Negative correlation	one element of paired scores increases while the other decreases
Pearson's r	measure of association between **two interval\ratio variables**
Phi Coefficient	measure of association between **two dichotomous variables**
Point biserial	measure of assoc'n between an **interval/ratio variable** and a dichotomous variable
Positive correlation	both elements in paired scores increase (or decrease)
Rank biserial	measure of assoc'n between **ordinal variable** and dichotmous variable
Scattergram	**graphical representation** of correlation of two variables
Spearman's rho	measure of association between **two sets of ranks** (n > 10 pairs)

Study Questions

1. Is age related to the length of stay of surgical patients in a hospital? The following data was collected in a recent study.

Age:	40	36	30	27	24	22	20
Days:	11	9	10	5	12	4	7

a. Draw a scatterplot diagram of the data, with AGE on x-axis and DAYS on y-axis.
b. By appearance alone, do AGE and DAYS appear to be related?
c. Compute the appropriate correlation coefficient.
d. Interpret the results.
e. Compute the coefficient of determination. What does it tell you?

2. A professional person and a blue-collar worker were asked to rank 12 occupations according to the social status they attached to each. A ranking of 1 was assigned to the occupation with the highest status down to a ranking of 12 for the lowest. Here are their rankings:

Occupation	Professional Person	Blue-Collar Worker
Physician	1	1
Dentist	4	2
Attorney	2	4
Pharmacist	6	5
Optometrist	12	9
School Teacher	8	12
Veterinarian	10	6
College Professor	3	3
Engineer	5	7
Accountant	7	8
Health Care Administrator	9	11
Government administrator	11	10

a. Compute the appropriate correlation coefficient.
b. Interpret the results.

Sample Test Question

Match the following statistical procedures with the types of kinds of data below.

a. Pearson's r b. Spearman's rho c. Phi (r_ϕ)

d. Point Biserial e. Rank Biserial f. Cramer's Phi (Chap 23)

____ 1. Preacher popularity by rank and whether he graduated from seminary or not.

____ 2. Reading score in 6-year-olds and whether they participated in HEADSTART preschool program.

____ 3. Seminary GPA and marital satisfaction scores of graduating students.

____ 4. Smoking/not smoking and death by (1) cancer or (2) "other causes."

____ 5. Staff position and leadership style category.

____ 6. Ten objectives in Christian Education ranked by pastors and ministers of education.