Correlation: Measure of Relationship

• Bivariate Correlations are correlations between two variables. Some bivariate correlations are nondirectional and these are called symmetric correlations. Other bivariate correlations are directional and are called asymmetric correlations.

• Bivariate correlations control for neither antecedent variables (previous) nor intervening (mediating) variables.
Example 1: An antecedent variable may cause both of the other variables to change. Without the antecedent variable being operational, the two observed variables, which appear to correlate, may not do so at all. Therefore, it is important to control for the effects of antecedent variables before inferring causation.
Example 2: An intervening variable can also produce an apparent relationship between two observed variables, such that if the intervening variable were absent, the observed relationship would not be apparent.

• The linear model assumes that the relations between two variables can be summarized by a straight line.
• Correlation means the co-relation, or the degree to which two variables go together, or technically, how those two variables covary.
• Measure of the strength of an association between 2 scores.
• A correlation can tell us the direction and strength of a relationship between 2 scores.
• The range of a correlation is from –1 to +1.
• -1 = an exact negative relationship between score A and score B (high scores on one measure and low scores on another measure).
• +1 = an exact positive relationship between score A and score B (high scores on one measure and high scores on another measure).
• 0 = no linear association between score A and score B.

• When the correlation is positive, the variables tend to go together in the same manner.
Example: As a person's score on one variable goes up, their score on the second variable also tends to go up. If someone scores a low score on one variable, we would expect them to also score low on the second variable.

• When the correlation is negative, we tend to see an inverse or opposite direction in the relationship.

Example: As a person's score on one variable goes up, their score on the second variable would tend to be lower. If someone scores a low score on the first variable, we would actually expect them to now score higher on the second variable.

• **Partial Correlation:** Shows relationship between x and y while holding z constant. This correlation is applied to control for potentially confounding variables in correlation analysis.

Correlation Indices

Type of Correlation	Symbol	Types of variables	Example
Pearson's *r*	r	2 continuous variables	Height and weight
Spearman rho	ρ or r_s	At least one variable is ordinal level	Placement of finish in a race (ordinal level) and muscle mass
Biserial *r*	r_b	Both variables are continuous but one has been arbitrarily dichotomized	Score on employment test (continuous) with rating of "satisfactory" or "unsatisfactory" in terms of numbers of errors made on job (arbitrary dichotomy)
Point biserial *r*	r_{pb}	Correlation between one continuous variable and a variable that is a true dichotomy	Correlation between height (continuous) and gender (true dichotomy)
Tetrachoric	r_t	Correlation between two continuous variables that have been arbitrarily dichotomized	-Correlation of tall versus short (arbitrarily dichotomized) with pass versus fail a physical fitness test (arbitrarily dichotomized) -Correlation between pass/fail an entrance exam and good/poor student
Phi	ϕ	Correlation between two true dichotomous variables.	Correlation between male/female and alive/dead.

Coefficient (*r*)	Correlation	Interpretation
$r < .20$	slight correlation	almost no relationship
r .21 to .40	low correlation	small relationship
r .41 to .70	moderate correlation	substantial relationship
r .71 to .89	high correlation	distinct relationship
$r > .90$	very high correlation	solid relationship

I. Interval Variables

Pearson's *r* or Pearson's Product-Moment Correlation Coefficient

• This indicates the percentage of the strength of the relationship between 2 sets of scores.

Assumptions:

1. Interval level data.
2. The variables being correlated must be paired observations.
3. Linearity: Plot the relationship between the variables with a scatterplot or fit the functional curve formed by the relationship to be sure of linearity and not curvilinearity.
4. Bivariate normality.
5. Homoscedasticity or equal variances: Truncated variances can attenuate the relationship.
6. Independence of observations.
7. Representative sampling.

1. H_0: $p = 0$ (p is rho)
2. H_1: p not equal to 0 or also $p < 0$ (negative correlation); $p > 0$ (positive correlation)

Computational Formula

$$r = \frac{\sum xy / N - (M_x)(M_y)}{SD_x SD_y}$$

Example:

Student	Hrs. Study **x**	$\mathbf{x^2}$	GPA **y**	$\mathbf{y^2}$	**xy**
A	40	1600	3.75	14.063	150.000
B	30	900	3.00	9.000	90.000
C	35	1225	3.25	10.563	113.750
D	5	25	1.75	3.063	8.750
E	10	100	2.00	4.000	20.000
F	15	225	2.25	5.063	33.750
G	25	625	3.00	9.000	75.000
	= 160	= 4700	= 19.000	= 54.752	= 491.250

1. $M_x = \sum x / N = 160 / 7 = 22.857$

2. $M_y = \sum y / N = 19.00 / 7 = 2.714$

3. $SD_x = \sqrt{4700 / 7 - 22.857^2} = 12.206$

4. $SD_y = \sqrt{54.752 / 7 - 2.714^2} = .675$

5. $r = \frac{\sum xy / N - (M_x)(M_y)}{SD_x SD_y} = \frac{491.250 / 7 - 62.034}{(12.206)(.675)} = \frac{8.145}{8.239} =$ **.989**

Conclusion: Looking at the table for the critical value of a two-tailed test with n = 7, df = 5, and alpha = .05, we find a critical value = .754. Thus, we reject the H_0 at the .05 level because our obtained value of .989 is greater than .754. We can say that there is a statistically significant correlation in the population and we have a very strong, positive relationship between hours studied and GPA.

Note: An *r* = .989 can be squared or .978
This r^2 is called the coefficient of determination and tells us the proportion of the total variance in Y that can be associated with the variance in X.
Thus, about 98% of the variance in GPA can be associated with the variance in hours studied.

II. Ordinal Variables

• There are correlations that are applied to two ordinal kinds of variables. These are typically nonparametric correlations. These correlation coefficients are distribution free and are usually applied to the ranks of the two variables.
• They measure monotonicity or whether one variables changes in the same direction as the other variable, when changes from one case to the next is considered.
• If both variables change in the same direction, a concordance is found.
• If one variable changes in one direction while the other variable changes in the opposite direction, a discordance is found.
• The total number of concordances and the total number of discordances for all pairs of observations are counted.

1. Spearman's Rank-Order Correlation (r_s)

Example:
Two judges rate the performance of 10 students on a particular skill.

$$r_s = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

Students	X	Y	$D = (X_1 - Y_2)$	D^2
1	2	3	-1	1
2	3	1	2	4
3	7	5	2	4
4	6	9	-3	9
5	1	2	-1	1
6	5	6	-1	1
7	10	8	2	4
8	8	10	-2	4
9	9	7	2	4
10	4	4	0	0
				$\sum d^2 = 32$

X = Judge 1
Y = Judge 2
D = The difference between the ratings of Judge 1 and Judge 2

$= 1 - \frac{6(32)}{10(100 - 1)} = \frac{192}{10(99) \text{ or } 990}$

$= 192 / 990 = .1939$

$= 1 - .194 =$ **.806 or .81**

Conclusion: Looking at the table for the critical value of Spearman's correlation with n = 10 and alpha = .05, we find a critical value = .649. Thus, we reject the H_0 at the .05 level because our obtained value of .806 is beyond the critical region of .649. We can say that there is a statistically significant correlation in the population and there is a strong, positive relationship between the two judges pertaining to their ranking of the students.

2. Kendall's Tau (τ)

• Like Spearman's, τ is a rank correlation method, which is used with ordinal data.
• The value of τ goes from –1 to +1.
• Tau is usually used when N < 10.

Formula:

$$\tau = \frac{C-D}{.5N(N-1)}$$

C = The number of pairs that are concordant or ranked the same on **Both** X and Y
D = The number of pairs that are discordant or inverted ranks on X and Y
Example:
We have two sets of ranks of Fred (X) and Sally (Y) on an intelligence measure:
Sample 1 (X): 1 2 3 4 5
Sample 2 (Y): 1 4 3 5 2
Note: The X ranks are in their natural order and the Y ranks exhibit a degree of disarray.

1. If a pair is ranked in its natural order, such as 1 and 4, a weight of + is assigned. If a pair is ranked in an inverse order, such as 4 and 3, a weight of – is assigned.
2. **Sample 1**: + + + + -
   Sample 2: + - + - -
3. C = 6 and D = 4 or 6-4 = 2
4. .5x5 = 2.5x4 = 10
5. 2 / 10 = .20 **τ = .200**
6. **Note:** The more concordant pairs than discordant, produces a positive relationship of X and Y.

Conclusion: Looking at the table for the critical value of Kendall's correlation with n = 5 and alpha = .05, we find a critical value = .800. Thus, we fail to reject the H_0 at the .05 level because our obtained value of .200 is not beyond the critical region of .800. We can say that there is not a statistically significant correlation in the population and there is a weak, positive relationship between Fred and Sally's scores.

Question: So, if I know that Fred is ranked higher than Sally, does this help me make a prediction about their rank order on y?
Answer: In this instance, tau = .200, which is a weak relationship. This does not help much in terms of predictions about their rank order on Y.

III. Dichotomous Variables

1. Phi

• The kind of correlation that is applied to two binary variables is the phi correlation.
• Special case of Pearson's r when both variables are dichotomous (see crosstabulation table).

	Gender - nominal: Male (0)	Female (1)
Republican (1)	2 A	4 B
Democrat (0)	3 C	1 D

$$\phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

$$= \frac{(4)(3)-(2)(1)}{\sqrt{(6)(4)(5)(5)}} = \frac{12-2}{\sqrt{600}}$$

product of the "marginals" $= \frac{10}{24.5} =$ **.408**

2. Point-Biserial Correlation Coefficient (r_{pb})

Special case of Pearson's r when one variable is interval/ratio and other variable is dichotomous

*	Group 1 = T 0 = C		Test Score				
Subject	(x)		(y)		x^2	y^2	xy
A	1	correct	10		1	100	10
B	1		12		1	144	12
C	1	p=5/10=.5	16		1	256	16
D	1		10	$\bar{y}_1 = 11.80$	1	100	10
E	1		11		1	121	11
F	0		7		0	49	0
G	0	incorrect	6		0	36	0
H	0		11	$\bar{y}_o = 7.40$	0	121	0
I	0	q=5/10=.5	8		0	64	0
J	0		5		0	25	0
	$\Sigma x = 5$		$\Sigma y = 96$		$\Sigma x^2 = 5$	1,016	59
						$= \Sigma y^2$	$= \Sigma xy$

(n = 10)

*the larger the sample the more normal the curve

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{\sigma_y}\sqrt{pq} = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{\left[\Sigma y^2 - (\Sigma y)^{2/n}\right]/n}}\sqrt{pq} = \frac{11.80 - 7.40}{\sqrt{\left[1{,}016 - (96)^{2/10}\right]/10}}\sqrt{(.5)(.5)}$$

$= \frac{4.40}{3.07}\sqrt{(.5)(.5)} =$ **.716** (same as Pearsonian *r* calculation, simplified because one variable, x, is a dichotomy).

Conclusion: Looking at the table for the critical value of a two-tailed test with n = 10, df = 8, and alpha = .05, we find a critical value = .632. Thus, we reject the H_0 at the .05 level because our obtained value of .716 is greater than .632. We can say that there is a statistically significant correlation in the population and we have a very strong, positive relationship between test scores and the treatment group.

2B. Point-Biserial Correlation Coefficient (r_{pb}) Found via an Independent Samples *t*-Test

Test Statistic $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$\mu = 11.80$	$\mu = 7.40$
$s = 2.49$	$s = 2.30$
$n = 10$	$n = 10$

Finding the *t* statistic:
Numerator:
11.80 – 7.40 = 4.40

Denominator:
$2.49^2 / 5 = 1.24$
$2.30^2 / 5 = 1.06$
$1.24 + 1.06 = 2.30^{.5} = 1.52$

Final Step for t:
4.40 / 1.52 = **2.895 is the t value**

Conversion Step to r_{pb}
SQRT (r^2) = t^2 / df + t^2
$2.895^2 / 8 + 2.895^2 = 8.381 / 16.381 = .512$

= SQRT (.512) = **.716 or the same Point-Biserial Correlation Coefficient**

Conclusion: Looking at the table for the critical value of a two-tailed test with n = 10, df = 8, and alpha = .05, we find a critical value = .632. Thus, we reject the H_0 at

the .05 level because our obtained value of .716 is greater than .632. We can say that there is a statistically significant correlation in the population and we have a very strong, positive relationship between test scores and the treatment group.