ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/277324930

Pearson's Product-Moment Correlation: Sample Analysis

Research · May 2015

DOI: 10.13140/RG.2.1.1856.2726

CITATIONS
0

READS
29,907

1 author:

Jennifer D. Chee
The Queen's Medical Center
19 PUBLICATIONS **23** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

NURS 220L Clinical Course Lecturer View project

2016 Queen's Health System Educational Priorities Survey View project

All content following this page was uploaded by Jennifer D. Chee on 29 May 2015.

The user has requested enhancement of the downloaded file.

Running head: PEARSON'S PRODUCT MOMENT CORRELATION

Pearson's Product Moment Correlation: Sample Analysis

Jennifer Chee

University of Hawaii at Mānoa School of Nursing

Introduction

Pearson's product moment correlation coefficient, or *Pearson's r* was developed by Karl Pearson (1948) from a related idea introduced by Sir Francis Galton in the late 1800's. In addition to being the first of the correlational measures to be developed, it is also the most commonly used measure of association. All subsequent correlation measures have been developed from Pearson's equation and are adaptations engineered to control for violations of the assumptions that must be met in order to use Pearson's equation (Burns & Grove, 2005; Polit & Beck, 2006). Pearson's *r* measures the strength, direction and probability of the linear association between two interval or ratio variables.

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

Pearson's *r* is a measure of the linear relationship between two interval or ratio variables, and can have a value between -1 and 1. It is the same measure as the point-biserial correlation; a measure of the relationship between a dichotomous (yes or no, male or female) and an interval/ratio variable (Cramer, 1998).

The advantage of using Pearson's *r* is that it is a simple way to assess the association between two variables; whether they share variance (*covary*), if the relationship is positive or negative, and the degree to which they correlate. The disadvantages of using Pearson's r is that it can not identify relationships that are not linear, and may show a correlation of zero when the correlation has a relationship other than a linear one. Additionally the types of variable that can be evaluated are limited.

In addition to Pearson's *r*, *semipartial* and *partial correlation* can be employed in order to estimate the relationship between an outcome and predictor variable after controlling for the effects of additional predictors in the equation.

. "Pearson's correlation is the ratio of the variance shared by two variables" (Cramer, 1998, p. 137). A means of illustrating correlation is with a Venn diagram (Figure 1). Each circle represents the amount of variance of each variable. In Figure 1a the circles do not overlap, indicating that there is no correlation between the two variables. In Figure 1b the two circles overlap, the size of the correlation is reflected in the degree of overlapping area. In a perfect correlation (Figure 1c) the circles would overlap each other completely.

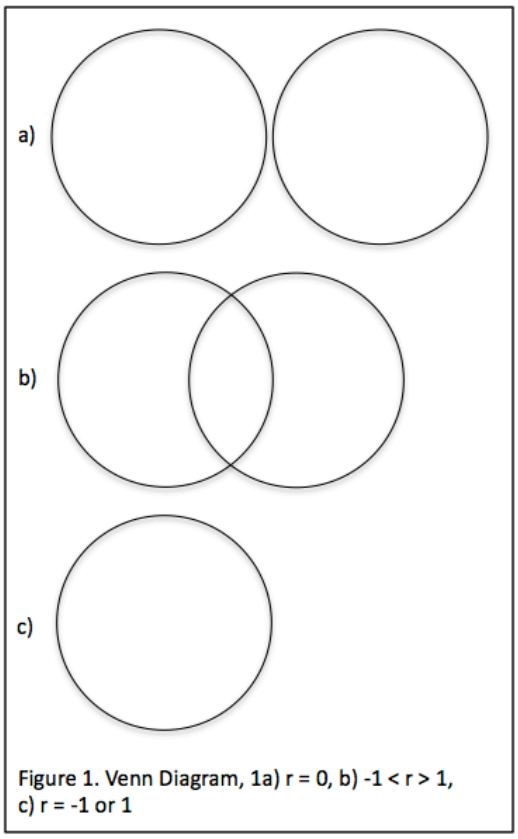


Figure 1. Venn Diagram, 1a) r = 0, b) -1 < r > 1, c) r = -1 or 1

There are multiple formulae that can be used to compute Pearson's *r*:

For small samples, Pearson's r may be calculated manually using:

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

Where:

r = Pearson's correlation coefficient

n = number of paired scores

X = score of the first variable

Y = score of the second variable

XY = the product of the two paired scores

Or similarly:

$$r = \frac{\text{covariance of variable } A \text{ and } B}{\sqrt{(\text{variance of variable } A) \times (\text{variance of variable } B)}}$$

Analogously:

$$r = \frac{\text{covariance of variable } A \text{ and } B}{(standard\ deviation\ of\ variable\ A) \times (standard\ deviation\ of\ variable\ B)}$$

"The larger the covariance is, the stronger the relationship. The covariance can never be bigger that the product of the standard deviation of the two variables" (Cramer, 1998, p. 139).

The denominator in the equations is always positive, however the numerator may be positive, zero, or negative, "…enabling *r* to be positive, negative, or zero respectively" (Zar, 1999, p. 378).

Assumptions

Of Primary importance are *linearity* and *normality*. Pearson's *r* requires that interval data be used to determine a linear relationship. Further assumptions must be met in order to establish statistical significance, "…for the test statistic to be valid the sample distribution has to be normally distributed" (Filed, 2009, p. 177). Homoscedasticity assumes that the error at each level of the independent variable is constant. A violation of the assumption of homoscedasticity increases the chances of obtaining a statistically significant result even though H_0 is true.

Significant Testing of Pearson's *r*

In order to infer that the calculated *r* is applicable to the population from which the sample was drawn, statistical analysis must be performed to determine whether the coefficient is significantly different from zero.

The hypothesis that the population correlation coefficient is 0, and may be computed by calculating *t*:

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

Where:

r = Pearson's product-moment correlation coefficient

n = sample size of paired scores

df = *n* - 2

If the sample size is small, a high correlation coefficient (close to -1 or 1), may be non-significant. Contrariwise, a large sample may have a statistically significant *r* but have no clinical significance. Subsequently it is important to consider both the magnitude of the correlation coefficient, the significance of the t-test, and the context of the research question.

Coefficient of Determination

By squaring the correlation coefficient *r*, the total variability in Y can be accounted for after regressing Y on X; r^2 can be considered to be a measure of the strength of the linear relationship. The resulting value when multiplied by 100 results in a percent variance, e.g., if the correlation coefficient for X and Y is $r = .50$, then $r^2 = (.50)(.50) = .25 = .25(100) = 25\%$. X explains 25% of the variability in Y (Zar, 1999).

The Matrix

The output from a statistical program such as SPSS contains a single table, the *correlation matrix*. The correlation analysis computes the correlation coefficient of a pair of

variables - or if the analysis contains multiple variables - the matrix contains the results for all variables under consideration, and is obtained by performing bivariate correlational analysis on every possible pairwise combination of variables in the dataset. "On the matrix, the *r* value and the *p* value are given for each pair of variables. At a diagonal through the matrix are variables correlated with themselves" (Burns & Grove, 2005, p. 488), the value of *r* for these correlations is equal to 1 and the *p* value is .000 because a variable is perfectly related to itself.

Pearson's Partial Correlation

A statistically significant association between two measures does not mean that the two variables have a causal relationship, e.g. if students who spend more time in the simulation laboratory were found to have higher scores during skills evaluations, this association does not necessarily imply a causal link between *hours in the 'sim lab'* and *higher marks on evaluations*, it is conceivable that the relationship is specious (Zar, 1999). Such relationships are referred to as *spurious*. Such spurious associations are the result of one or more other factors that are "genuinely related to these two variables" (Cramer, 1998, p. 156). To continue the example, a student's *curricular progression* may be related to both hours in the 'sim lab' and higher marks on evaluations. When curricular progression is taken into consideration there may be *no* association between the other two variables. A more likely example of a spurious relationship is one in which *part but not all* of the association between two measures may be due to one or more other factors, e.g., students who spend more time in the simulation laboratory may have had more exposure to nursing skills secondary to their progression through the curriculum. If the influence of the student's position within the curriculum on the association between hours in 'sim lab' and higher marks on evaluations is partialled out, it can be presumed that the remaining association between these two variables is not a direct result of position within the curriculum.

Similarly, removing the influence of hours in 'sim lab' from the relationship between position within the curriculum and higher marks on evaluations to determine how much of the relationship between the two variables was not due to hours in 'sim lab'.

> It is important to note that a significant reduction in the size of an association between two variables by partialling out the influence of a third variable does not necessarily mean that that reduction indicates the degree of spuriousness in the original relationship between the two variables. (Cramer, 1998, p. 157)

To perform a partial correlation to control for one variable – *first-order partial correlation* ($r_{12.3}$) the following formula may be used:

$$r_{12.3} = \frac{r_{12} - (r_{13} \times r_{23})}{\sqrt{(1 - 2_{13}^2) \times (1 - 2_{23}^2)}}$$

To perform a partial correlation to control for two variables – *second-order partial correlation* ($r_{12.34}$) the following formula, which is based on the formula for first order correlations, may be used:

$$r_{12.3} = \frac{r_{12.3} - (r_{14.3} \times r_{24.3)}}{\sqrt{1 - 2_{14.3}^2) \times (1 - 2_{24.3}^2)}}$$

And third-order partial correlations are similarly "based on second-order partial correlations and so on" (Cramer, 1998, p. 160; Pedhazur, 1982).

Pearson's Semipartial Correlation

Whereas partial correlation allows for partialling out a variable from two other variables, *semipartial correlation* (or part correlation) allows for the partialling out of a variable from only one the variables that are being correlated. Put another way, semipartial correlation is the correlation between the outcome variable and another predictor variable that has been controlled for by another predictor variable. For example assume an admissions committee at a school of

nursing is considering three variables, high school grade point average (*HS_GPA*), score on the Test of Essential Academic Skills (*TEAS*), and intelligence quotient (*IQ*). It would not be unreasonable to assume IQ and TEAS are positively correlated. In this situation the admissions committee is interested in the relationship between TEAS and HS_GPA, while controlling for IQ. The formula for $r_{12.3}$ (correlation of HS_GPA and TEAS while controlling for IQ) could provide some information. Also $r_{13.2}$ will indicate the correlation for IQ and HS_GPA while controlling for TEAS.

> It is possible however, that of greater interest to the [admissions committee] is the predictive power of the [TEAS] after that of [IQ] has already been taken into account…the interest is in the increment in the proportion of the variance in [HS_GPA] accounted for by the [TEAS] that is over and above the proportion of variance accounted for by [IQ]. In such a situation [IQ] should be partialled out from [TEAS], but not from [HS_GPA] where it belongs [by performing a semipartial correlation.] (Pedhazur, 1982, p. 115)

Sample Problem

The flexibility permitted by simulation has made it attractive to institutions as a supplement to real-world clinical experience (Murray, Grant, Howarth, & Leigh, 2008; Rosseter, 2011). Resource constraints by academic institutions, fiscal reforms and societal pressures have promoted a safety-conscious-litigation-avoiding culture where simulation provides a means of risk free applied practice (Health care at the crossroads: Strategies for improving the medical liability system and preventing patient injury, 2005).

Effective use of simulation is dependent on a complete understanding of the role of simulation's central conceptual elements. *Deliberate practice*, a constituent of Ericsson's theory

of expertise, has been identified as a central concept in effective simulation learning (Issenberg, McGaghie, Petrusa, Gordon, & Scalese, 2005).

Deliberate practice is compatible with simulation frameworks already being suggested for use in nursing education (Jeffries, 2006). In a review of the websites of leading nursing schools in the United States ("Top Nursing Schools and Best Ranked Nursing Colleges," 2011), each mentioned access to simulation centers or labs. In an informal non-scientific survey of these same schools, respondents reported use of low, moderate, and high fidelity simulation in addition to task trainers and moulage/tabletop models. The respondents reported that, depending on the students' class standing, they are required to complete anywhere from 100 – 280 real-world clinical hours, and a minimum of 4 simulated clinical hours (.04 and .014% respectively) in the simulation lab during adult health medical/surgical semesters. The same institutions describe simulation as a *compliment to* rather than a *replacement for* clinical experience (Chee, 2011).

A small undergraduate-nursing program was considering a substantial increase in the amount of curricular time the nursing students spent in the simulation laboratory ('sim lab'). Before making the changes the school wished to evaluate the impact of simulation on the biannual skills evaluations. The 'sim lab' was made available to the students on a voluntary basis. The 'sim lab' was equipped with evidence-based multimedia learning modules for the purpose of independent deliberate practice of clinical nursing skills.

Data was collected regarding the number of hours each of the undergraduate-nursing students spent in the simulation laboratory participating in deliberate practice, pre- and posttest evaluation scores, and some demographic information were collected. The course instructors hypothesized that there was no difference in the evaluation scores depending on the amount of time students spent participating in deliberate practice for the purpose of learning nursing skills.

SPSS 21 was used to perform all assumption testing and analysis.

Null and Alternative Hypothesis

The null and alternative hypothesis for the correlation is:

H_0: $\varrho = 0$, the population correlation coefficient is equal to zero

And the alternative hypothesis is:

H_A: $\varrho \neq 0$, the population correlation coefficient is not equal to zero.

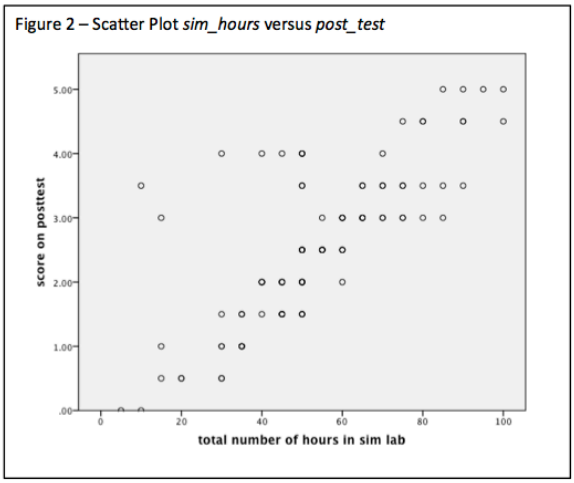
Where ϱ is the population correlation coefficient

Results

Testing of Assumptions

A Pearson correlation is appropriate only when a linear relationship exists between the two variables. To determine if a linear relationship existed between the primary variables of interest a scatter plot (Figure 2) was generated between the predictor (*sim_hours*) and the outcome variable (*post_test*). The relationship approximated a straight line; the scatterplot suggests that the relationship of the two variables is positive and linear; the assumption of linearity has not been violated.

Figure 2 – Scatter Plot *sim_hours* versus *post_test*



Next the assumption of bivariate normality was evaluated. Both variables were tested for normality using the Shapiro-Wilk test (Table 1), which was appropriate for the sample size. Both variables are normally distributed, as assessed by Shapiro-Wilk's test ($p > .05$).

Table 1 – Shapiro Wilk Test of Normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
sim_hours	.115	100	.002	.981	100	.169
post_test	.087	100	.059	.976	100	.059

a. Lilliefors Significance Correction

Pearson's Product Moment Correlation Coefficient

The main analysis was performed with *sim_hours* as the predictor variable and *post_test* as the outcome variable (Table 2). There was a positive correlation between hours spent practicing in the 'sim lab' and post-test scores, $r(98) = .751$, $p < .0005$. The relationship was statistically significant; therefore we reject the null hypothesis and accept the alternative hypothesis; there is a correlation between time spent in the 'sim lab' and the successful execution of psychomotor skills in a posttest evaluation.

Table 2 – Correlation of *sim_hours* versus *post_test*

Correlations

		number of hours in sim lab	score on posttest
sim_hours	Pearson Correlation	1	.751**
	Sig. (2-tailed)		.000
	N	100	100
post_test	Pearson Correlation	.751**	1
	Sig. (2-tailed)	.000	
	N	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Strength of Association

The absolute value of the Pearson coefficient determines the strength of the correlation. Using Cohen's (1988) guidelines for interpreting strength of association ($.1 < |r| < .3$ a small correlation, $.3 < |r| < .5$, a moderate correlation, $|r| > .5$ a strong correlation), the results of this analysis show a *strong positive correlation* between the predictor and outcome variables.

Coefficient of Determination

Hours spent in the 'sim lab' explained 60% of the variability in post-test scores: r^2 = (.751)(.751) = .564 = .564(100) = 56%.

Partial Correlation

The instructors suspected that the correlation between time spent in the 'sim lab' and posttest evaluation scores was only a spurious correlation caused by a difference in the baseline characteristics of the students. Subsequently they investigated the relationship of the variables *sim_hours* and *post_test* further by evaluating the influence of other data that had been collected.

A second correlational analysis was performed (Table 3) which included a second predictor variable, *pre_test*, suspected of having a possible influence on *post_test*.

From this matrix (Table 3) it was known that *post-test* scores are strongly positively correlated with hours spent in the *sim_lab*, and it was also known that *post_test* scores were highly positively correlated with *pre_test*. But what was unknown was what the correlation would be when *pre_test* was controlled for.

Table 3 – Correlation Matrix

Correlations

		score on posttest	new prestest scores	number of hours in sim lab
post_test	Pearson Correlation	1	.927**	.751**
	Sig. (2-tailed)		.000	.000
	N	100	100	100
pre_test	Pearson Correlation	.927**	1	.665**
	Sig. (2-tailed)	.000		.000
	N	100	100	100
sim_hours	Pearson Correlation	.751**	.665**	1
	Sig. (2-tailed)	.000	.000	
	N	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

The evaluation was accomplished with a partial correlation. The output (Table 4) displayed the zero order correlation coefficients, which were the three Pearson's correlation coefficients without any control variable taken into account. The zero-order correlations supported the hypothesis that more hours spent in the 'sim lab' increased the post-test score. A strong association of r = .751, which was highly significant. The variable *pre_test* was also significantly correlated with both *sim_hours* and *post_test* score (r = .665 and r = .927). The second part of the output showed the correlation coefficient between *sim_hours* and *post_test* when *pre_test* was controlled for. The correlation coefficient was now, $r_{12.3} = .479$, and was statistically significant, $p < .005$.

Table 4 – Partial Correlation Matrix

Correlations

Control Variables			number of hours in sim lab	score on posttest	new prestest scores
-none-[a]	sim_hours	Correlation	1.000	.751	.665
		Significance (2-tailed)	.	.000	.000
		df	0	98	98
	post_test	Correlation	.751	1.000	.927
		Significance (2-tailed)	.000	.	.000
		df	98	0	98
	pre_test	Correlation	.665	.927	1.000
		Significance (2-tailed)	.000	.000	.
		df	98	98	0
pre_test	sim_hours	Correlation	1.000	.479	
		Significance (2-tailed)	.	.000	
		df	0	97	
	post_test	Correlation	.479	1.000	
		Significance (2-tailed)	.000	.	
		df	97	0	

a. Cells contain zero-order (Pearson) correlations.

Thus when the influence of the variable *pre_test* was controlled for the *sim_lab* accounted for only 23% of the variability in *post_test*.

Discussion

The second (partial) correlation was useful in understanding the impact of factors other than *sim_hours* on students' nursing skill evaluations. The instructors should continue to look at elements of student experience and curriculum that contribute to pre- and posttest scores in an effort to make intelligent decisions about how best to focus resources.

Conclusion

Pearson's product moment correlation coefficient (Pearson's r) is not conceptually demanding; which underlies it's utility. It is one of the foundational elements of modern statistical analysis, and is a simple means of evaluating the linear relationships between variables. The direction (positive or negative) and strength of the relationship (coefficient of determination) may also be evaluated. While it does not show causal relationships it is an pragmatic way to understand the relationship of variables of interest.

References

Burns, N., & Grove, S. (2005). *The Practice of Nursing Research: Conduct, Critique, and Utilization* (5 ed.). St. Louis: Elsevier Saunders.

Chee, J. (2011). [Simulation Use in Undergraduate Nursing Education].

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Academic.

Cramer, D. (1998). *Fundamental Statistics for Social Research*. London: Routledge.

Crown, W. (1998). *Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-Dependent Variable Models*. London: Praeger.

Issenberg, S., McGaghie, W., Petrusa, E., Gordon, D., & Scalese, R. (2005). Features and uses of high-fidelty medical simulations that lead to effective learning: A BEME systematic review. . *Medical Teacher, 27*(1), 10-28. doi: 10.1080/01421590500046924

Jeffries, P. (2006). Developing evidenced-based teaching strategies and practices when using simulations. *Clinical Simulation in Nursing, 2*(1), e1-e2. doi: 10.1016/j.ecns.2009.05.014

Murray, C., Grant, M., Howarth, M., & Leigh, J. (2008). The use of simulatoin as a teaching and learning approach to support practice learning. *Nurse Education in Practice, 8*, 5-8.

Pearson, K. (1948). *Early Statistical Papers*. Cambridge, England: University Press.

Pedhazur, E. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction*. New York: Holt, Rinehart and Winston.

Polit, D., & Beck, C. (2006). *Essentials of Nursing Research: Methods, Appraisal, and Utilization* (6 ed.). Philadelphia: Lippincott Williams & Wilkins.

Rosseter, R. (2011). Nursing shortage fact Sheet: American Associaltion of Colleges of Nursing.

Zar, J. H. (1999). *Biostatistical Analysis*. Upper Saddle River, NJ: Prentice Hall.

View publication stats