On the Nonadditivity of Correlation Coefficients
Part 1: Pearson's *r* and Spearman's r_s

Abstract – This is Part 1 of a tutorial series on the nonadditivity of correlation coefficients. We demonstrate why it is not possible to arithmetically add, subtract, and average Pearson's *r* or Spearman's r_s.

Keywords: nonadditivity, correlations, coefficients, pearson, spearman
Published: 01-07-2011; Updated: 07-04-2017

Note: This article is part of a legacy series that the author published circa 2011 at http://www.miislita.com, now a search engine site. It is now republished in pdf format here at http://www.minerazzi.com, with its content edited and updated.

Introduction

The purpose of this article is to demonstrate the nonadditivity of correlation coefficients.

As there are many types of correlations, the discussion is limited to Pearson product-moment correlation coefficient, *r*, and Spearman rank-order correlation coefficient, r_s. The following conventions are used:

- k is the number of samples.
- n is the sample size or number of paired observations (x, y) in a sample.
- υ is the number of degrees of freedom, where $\upsilon = n - 1$.
- x is an independent variable, $\bar{x}$ its mean, and s_x its standard deviation.
- y is a dependent variable, $\bar{y}$ its mean, and s_y its standard deviation.
- d is the difference between any two paired variables; i.e., $d = x - y$.
- r is Pearson's product-moment correlation coefficient between the paired variables (x, y).
- r_s is Spearman's rank-order correlation coefficient between the paired variables (x, y).
- β_0 is the intercept of a simple linear regression model.
- β_1 is the slope of a simple linear regression model.

Discussion

Densities, standard deviations, sines, cosines, tangents, and slopes are not additive because in each case the results are not new densities, standard deviations, sines, cosines, tangents, and slopes. In general, dissimilar ratios and intensive properties are not additive. By contrast, lengths, areas, variances, covariances, and angles are additive because in each case the results are new lengths, areas, variances, covariances, and angles.

Quantities that are not additive cannot be subtracted or averaged in the arithmetic sense either. For instance, let $x_1 = \cos(\theta_1)$ and $x_2 = \cos(\theta_2)$. From the cosine addition rule,

$$\cos(\theta_1) + \cos(\theta_2) = 2\cos\left(\frac{\theta_1+\theta_2}{2}\right)\cos\left(\frac{\theta_1-\theta_2}{2}\right) \quad (1)$$

$$\frac{\cos(\theta_1)+\cos(\theta_2)}{2} = \cos\left(\frac{\theta_1+\theta_2}{2}\right)\cos\left(\frac{\theta_1-\theta_2}{2}\right) \quad (2)$$

$$\cos(\theta_1) - \cos(\theta_2) = -2\sin\left(\frac{\theta_1+\theta_2}{2}\right)\sin\left(\frac{\theta_1-\theta_2}{2}\right) \quad (3)$$

$$\frac{\cos(\theta_1)-\cos(\theta_2)}{2} = -\sin\left(\frac{\theta_1+\theta_2}{2}\right)\sin\left(\frac{\theta_1-\theta_2}{2}\right) \quad (4)$$

The results from (1) to (4) are not new cosines. This conclusion also holds for Pearson's Correlation Coefficient, *r*.

Pearson's *r* Defined

Pearson's *r* is a parametric or distribution-dependent statistic. It is defined as the ratio of the covariance (*cov*) between two variables, normalized by their standard deviations.

$$r = \frac{cov_{xy}}{s_x s_y} = \frac{\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\upsilon}}{\frac{\sqrt{\sum_i^n (x_i - \bar{x})^2}}{\sqrt{\upsilon}} \frac{\sqrt{\sum_i^n (y_i - \bar{y})^2}}{\sqrt{\upsilon}}} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2}\sqrt{\sum_i^n (y_i - \bar{y})^2}} \quad (5)$$

In the numerator of (5), averages are subtracted from raw scores, and the sum of cross-products accumulated. In the denominator, variable scales are adjusted to have equal units, relative to the

sample distribution and degrees of freedom in question (Rodgers & Nicewander, 1988). Thus, if we try to average over k number of samples each with their own statistics, degrees of freedom, and normalized scales

$$\bar{r} = \frac{1}{k}\left(\frac{cov_{x_1y_1}}{s_{x_1}s_{y_1}} + \frac{cov_{x_2y_2}}{s_{x_2}s_{y_2}} + \cdots \frac{cov_{x_ky_k}}{s_{x_k}s_{y_k}}\right) \quad (6)$$

we actually end up adding dissimilar ratios.

Obviously, correlations are nonadditive in nature, but where does this nature come from? Mean-centering the variables provides us with an incontrovertible answer.

Pearson's *r* from Mean-Centered Variables

Mean-centered variables are obtained by subtracting their arithmetic means

$$x_i' = x_i - \bar{x} \quad (7)$$

$$y_i' = y_i - \bar{y} \quad (8)$$

In the same way that mean-centering does nothing for moderated multiple regression (Echambadi & Hess, 2007), here it does nothing to *r*, but merely removes variable fluctuations around their means so the new means, $\bar{x'}$ and $\bar{y'}$, vanish.

Thus, applying (7) and (8) to (5),

$$r = \frac{\Sigma_i^n (x_i')(y_i')}{\sqrt{\Sigma_i^n (x_i')^2}\sqrt{\Sigma_i^n (y_i')^2}} \quad (9)$$

A closer look at (9) shows that this is the same expression that one will obtain from the cosine of the angle between any two vectors, with the numerator representing a dot product and the denominator a product of vector magnitudes.

Thus, mean-centering unveils Pearson's r latent meaning and nature: it is a cosine and as such a nonadditive quantity.

$$r = \cos(\theta) \tag{10}$$

The implications of this findings are many. For instance, in the context of Information Retrieval (IR), (9) and (10) define the resemblance measure known as the *cosine similarity* between two vectors. Accordingly, it is not possible to arithmetically add, subtract or average cosine similarities.

These findings are not limited to IR. Consider Hunter-Schmidt's meta-analysis model (Hunter & Schmidt, 2000; Schmidt, Oh, & Hayes; 2009). In this model average effect sizes are commonly expressed as a weighted mean r of the form $\bar{r} = \sum_j^k n_j r_j / \sum_j^k n_j$ where k is the number of r's. If a constant sample size is used the model returns an arithmetic mean value! Since correlations are not additive, this is a flaw in the model and the results can be challenged. Said flaw can be compounded if we are dealing with negative and positive correlations. As noted by Field (2003),

> "For example, imagine we tested the efficacy of a powder ('Stat-Whizz') that could magically make you good at statistics. A trial in the USA found an effect size of .45, a replication in Belgium found an effect size of 0, and a further replication in the UK yielded an effect size of -.45. If we assume that these studies had equal sample sizes and so were equally weighted in the meta-analysis, then the resulting average effect size would be 0—there would be a non significant effect. Readers of such a meta-analysis might conclude, therefore, that Stat-Whizz was an ineffective drug. Of course, this conclusion is wrong: the drug worked in the USA, didn't work in Belgium and had a negative effect in the UK. As such, the issue of interest is not so much the overall effect of the drug, but at what levels the drug works: the fact that the drug doesn't work on the English is of little interest to all of the Americans for whom the drug is effective!"

By measuring effect sizes as correlations and equally weighting these, in this example the weighted average was reduced to an arithmetic average that did not describe the efficacy of the drug. Although Field did not use that example to make a point about the nonadditivity of correlations, "but at what levels the drug works", he correctly presents additional reasons for not arbitrarily adding correlations. The undisputed fact remains that doing so, not only is a bad practice, is an invalid mathematical exercise.

Pearson's *r* from Simple Linear Regression Coefficients

As noted by Rodger & Nicewander (1988), if a study reports the results of a simple linear regression model with intercept β_0 and slope β_1 as

$$y = \beta_0 + \beta_1 x \tag{11}$$

a correlation coefficient can be calculated from the square root of the coefficient of determination of the model, r^2, or from the product between the regression slope, β_1, and the ratio of standard deviations,

$$r = \beta_1 \left(\frac{s_x}{s_y}\right) \tag{12}$$

where $r = \beta_1$ for $s_x = s_y$. If we try to average

$$\bar{r} = \frac{1}{k}\left[\beta_{11}\left(\frac{s_{x_1}}{s_{y_1}}\right) + \beta_{12}\left(\frac{s_{x_2}}{s_{y_2}}\right) + \cdots \beta_{1k}\left(\frac{s_{x_k}}{s_{y_k}}\right)\right] \tag{13}$$

and we end up averaging products from dissimilar regression equations.

Now that we have demonstrated the nonadditivity of Pearson's *r*, let's examine if the same holds for Spearman's r_s.

Spearman Rank-Order Correlation Coefficient, r_s

A Spearman rank-order correlation coefficient, r_s, is a non-parametric or distribution-free statistic. If there are no ties, r_s is computed as

$$r_s = 1 - \frac{6}{n\,(n^2 - 1)} \sum_i^n d_i^2 \tag{14}$$

where $d_i = x_i - y_i$. If there are ties, one can assign the average to each of the equal values. The presence of tied ranks tends to lower the $\sum_i^n d_i^2$ term, but this effect is small unless there are a large

number of ties. In said case, Kendall *tau-b* is recommended over r_s (Kendall, 1938). In any event, if we try to average Spearman r_s's

$$\overline{r_s} = \frac{1}{k}\left[k - \frac{6}{n\,(n^2 - 1)}\left(\sum_i^n d_{i1}^2 + \sum_i^n d_{i2}^2 + \cdots \sum_i^n d_{ik}^2\right)\right] \quad (15)$$

and we end up adding the square of rank differences from dissimilar samples. The presence of ties in the individual samples makes the additions even more questionable.

Computing Spearman Rank Correlation as a Pearson's *r*

If there are no ties, $s_x = s_y$, $r_s = r$, and r_s can be computed as a Pearson coefficient from ranked variables, and from either the coefficient of determination or slope, β_1, of the regression equation. This is shown in Figure 1 where the course preferences of two students are examined.

Courses	Student A Rank	Student B Rank	*d*	d^2
Chemistry	1	2	-1	1
Computers	2	1	1	1
Mathematics	3	4	-1	1
Statistics	4	3	1	1
Genetics	5	7	-2	4
History	6	6	0	0
Music	7	5	2	4
sum	28	28	0	12
std. dev.	$s_x = 2.1602$	$s_y = 2.1602$		
Spearman's r_s	0.7857			
Pearson's *r*	0.7857			
β_1	0.7857			
$\sqrt{r^2}$	0.7857			

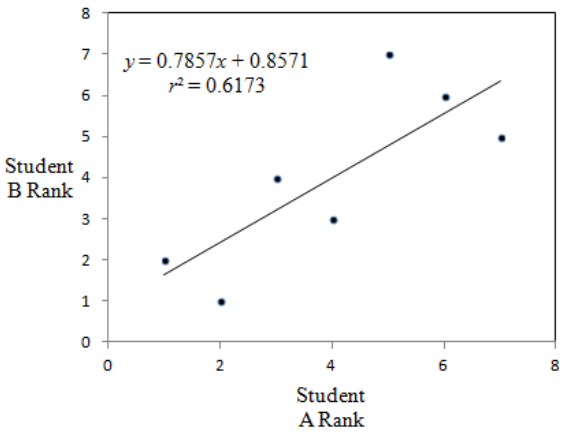

Figure 1. r_s computed as *r*.

In Figure 1, r_s was computed with (5), (12), and (14), and from the coefficient of determination, $r = \sqrt{r^2}$. Since computing the slope, β_1, of a straight line is the same as computing its tangent

$$slope = \frac{\Delta y_i}{\Delta x_i} = tan(\theta) \tag{16}$$

averaging over k numbers of r_s in this case is the same as averaging over several tangents,

$$\overline{r_s} = \frac{1}{k}(tan\,\theta_1 + tan\,\theta_2 + \cdots tan\,\theta_k) \tag{17}$$

But (17) is not possible because, like slopes, tangents are not additive; i.e.

$$\tan(\theta_1) + \tan(\theta_2) = \tan(\theta_1 + \theta_2)\,[1 - \tan(\theta_1)\tan(\theta_2)] \tag{18}$$

Thus, the sum of two or more tangents is not a new tangent. We must conclude that r_s's are not additive.

Conclusion

We have demonstrated why it is not possible to arithmetically add, subtract, or average Pearson or Spearman correlation coefficients.

Over the years several workarounds and weighting strategies have been proposed for reporting some forms of averages, most notoriously using Fisher's *Z* Transformations (Fisher, 1915; 1921). These are discussed in Part 2 of this series. The arbitrary implementation of these transformations will be refuted.

In the case of Spearman's r_s, we will show that these posses an inherent bias that depends of the population correlation and its sample size, *n*. This bias does not vanish even when *n* becomes infinite, increases when *n* decreases, and makes additions even more questionable.

References

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models, *Marketing Science* Vol. 26, No. 3, May–June 2007, 438-445. Retrieved from http://pubsonline.informs.org/doi/10.1287/mksc.1060.0263. See also http://www.bauer.uh.edu/jhess/papers/JMRMeanCenterPaper.pdf

Field, A. P. (2003). Can meta-analysis be trusted? The Psychologist, 16, 642-645. Retrieved from http://sro.sussex.ac.uk/714/1/CAN_META-ANALYSIS_BE_TRUSTED.pdf

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10,* 507-521. Retrieved from http://www.stat.duke.edu/courses/Spring05/sta215/lec/Fish1915.pdf

Fisher, R.A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3-32. Retrieved from http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15169/1/14.pdf

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: implications for cumulative research knowledge. International Journal of Selection and Assessment, 8, 275-292. Retrieved from https://www.researchgate.net/publication/279899123_Fixed_Effects_vs_Random_Effects_Meta-Analysis_Models_Implications_for_Cumulative_Research_Knowledge

Kendall, M. (1938). A new measure of rank correlation. Biometrika 30 (1-2): 81-89. Retrieved from https://www.jstor.org/stable/2332226?seq=1#page_scan_tab_contents

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, Vol. 42, No. 1, 59-66. Retrieved from https://www.stat.berkeley.edu/~rabbee/correlation.pdf

Schmidt, F. L., Oh, I., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. British Journal of Mathematical and Statistical Psychology, 62, 97-128. Retrieved from https://www.biz.uiowa.edu/faculty/fschmidt/meta-analysis/Schmidt_Oh_Hayes_2009.pdf