DigitalCommons@WayneState

Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 29

December 2017

JMASM 48: The Pearson Product-Moment Correlation Coefficient and Adjustment Indices: The Fisher Approximate Unbiased Estimator and the Olkin-Pratt Adjustment (SPSS)

David A. Walker
Northern Illinois University, dawalker@niu.edu

Recommended Citation

Walker, D. A. (2017). JMASM 48: The Pearson Product-Moment Correlation Coefficient and Adjustment Indices: The Fisher Approximate Unbiased Estimator and the Olkin-Pratt Adjustment (SPSS). Journal of Modern Applied Statistical Methods, 16(2), 540-546. doi: 10.22237/jmasm/1509496140

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 540-546.
doi: 10.22237/jmasm/1509496140

ISSN 1538 − 9472

JMASM 48: The Pearson Product-Moment Correlation Coefficient and Adjustment Indices: The Fisher Approximate Unbiased Estimator and the Olkin-Pratt Adjustment (SPSS)

David A. Walker
Northern Illinois University
DeKalb, IL

This syntax program is intended to provide an application, not readily available, for users in SPSS who are interested in the Pearson product–moment correlation coefficient (r) and r biased adjustment indices such as the Fisher Approximate Unbiased estimator and the Olkin and Pratt adjustment.

Keywords: Pearson product-moment correlation coefficient, SPSS, syntax

Introduction

The purpose for this computational program is to provide an application not readily available for users in the frequently employed Statistical Package for the Social Sciences (SPSS) software who are interested in the Pearson product-moment correlation coefficient (r) and r biased adjustment indices. The intent is that this program may assist users whose research importance is predicated on concepts such as point estimate bias or accuracy of point estimates to infer applicable and more robust suggestions about their data principally in a small sample size situation.

Correlation Coefficient

The Pearson product-moment correlation coefficient is employed extensively in social science research (Smithson, 2000) as a correlational technique between two variables (X and Y) and also in concurrence with numerous univariate and

David Walker is a Professor of Educational Research and Assessment. Email him at: dawalker@niu.edu.

multivariate methods "…to analyze the underlying relationship between the variables of interest prior to or following the main analysis" (Padilla & Veprinsky, 2014, p. 824). To be sure, there are alternative correlational methods that have been proposed to estimate the population correlation, ρ, (Donner & Rosner, 1980; Hotelling, 1953; Olkin, 1967), but Pearson's *r* appears to be the most frequently-applied statistic in this milieu.

Within the correlation coefficient's bivariate relationship between *X* and *Y*, it is assumed that this pairing has a linear relationship and both *X* and *Y* have a normal distribution (Olkin & Pratt, 1958), where "…observations follow a bivariate normal distribution with means (μ_{xi}, μ_{yi}), standard deviations (σ_{xi}, σ_{yi})…" (Donner & Rosner, 1980, p. 69). The sample correlation coefficient can be represented as

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}} \tag{1}$$

where *n* = number of *x*, *y* pairs, *xy* = product of *xy*, and Σxy = sum the product.

Fisher (1915, 1921, 1924) found that the correlation coefficient was comprised of an asymmetrical distribution, which also influenced this index's standard error, causing *r* to be a biased estimator of ρ under normal distribution conditions particularly with small sample sizes (i.e., for Fisher, "small" $N = 18$). Zimmerman, Zumbo, and Williams (2003) pointed out that the notion of "bias" in this situation is derived specifically from the sample mean associated with the *r* metric. Additionally, Zimmerman et al. noted that, practically,

> This discrepancy [bias] may not be crucial if one is simply investigating whether or not a correlation exists. However, if one is concerned with an accurate estimate of the magnitude of a non-zero correlation in test and measurement procedures, then the discrepancy may be of concern. (p. 134)

Bishara and Hittner (2015) established that the threshold for a "small" sample size was $N < 20$, where "… the absolute bias becomes negligible (less than .01) for a sample size greater than 20" (p. 786); noting that the bias decreased as the N increased. Further, Zimmerman et al. (2003) determined that the extent of the aforementioned estimation issue, where *r* could underestimate ρ by "…as much as .03 or .04 under some realistic conditions…" (p. 134). They also noted that *r* could achieve a positive bias as high as 0.05 under non-normal distribution conditions.

r-Based Bias Adjustments

To correct for the inherent bias affiliated with r, Fisher (1915) proposed the Fisher Approximate Unbiased (r_{FAU}) estimator, which assumes bivariate normality, and can be characterized as

$$r\left[1+\frac{\left(1-r^2\right)}{2n}\right] \tag{2}$$

where r = sample correlation coefficient. Additionally, Olkin and Pratt (1958), also assuming bivariate normality, suggested a second unbiased adjustment to r (r_{OP}), which can be represented as

$$r\left[1+\frac{\left(1-r^2\right)}{2\left(n-3\right)}\right] \tag{3}$$

Through a simulation study, Zimmerman et al. (2003) reported that the r_{OP} and the r_{FAU} adjustments were effectively the same when $N \geq 20$, but when $N < 20$, r_{OP} corrected bias more precisely than r_{FAU}. This finding was also corroborated in a simulation conducted by Walker (2016). Gorsuch and Lehmann (2010) supported the use of these r-based bias adjustments, though Bishara and Hittner (2015) were more cautious of their use in the presence of non-normal conditions.

Data and Programs

The example used here is comprised of a small sample, where N = 16, and are labor statistic data derived from Longley (1967). The full data set consists of seven economic-based variables measured from 1947 to 1962. The sample correlation is between the Y variable, the total derived employment, and an X variable, the number of people unemployed. As seen below, the user would enter in the program the sample correlation coefficient (r) and the sample size (N) in the space between BEGIN DATA and END DATA.

```
**********************************************************************
Copyright David A. Walker, 2016
Contact dawalker@niu.edu
Northern Illinois University, 325 Graham, DeKalb, IL 60115
             **APA 6th Edition Citation**
Walker, D. A. (2016). r and r biased adjustment indices [Computer program].
DeKalb, IL: Author.
**********************************************************************.

DATA LIST LIST /r (F8.3) N (F8.0).

**********************************************************************
NOTE: Between BEGIN DATA and END DATA, put the Pearson's Correlation
Coefficient (r) and the sample size (N)
**********************************************************************.

BEGIN DATA
.502 16
END DATA.
COMPUTE rFAU = ((1+(1-r**2)/(2*N))*r).
COMPUTE rOP = ((1+(1-r**2)/(2*(N-3)))*r).
COMPUTE FISHERZ = .5*LN((1+r)/(1-r)).
COMPUTE t = r*SQRT(N-2)/SQRT(1-r**2).
COMPUTE p1 = CDF.T(t,N-2).
COMPUTE p = (1-p1)*2.
COMPUTE Power = (1-CDFNORM(1.96-ABS(FISHERZ*SQRT(N-3)))).
COMPUTE r2 = r**2.
FORMAT rFAU TO r2 (F9.3).
VARIABLE LABELS r 'Pearson Correlation Coefficient r'/r2 'Variance
Explained by the Relationship r2'/ Power 'Post-Hoc Power'/p 'p-value'/rFAU
'Fisher Approximately Unbiased (rFAU) r'/rOP 'Olkin & Pratt (rOP) Adjusted
r'/.
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
   /VARIABELS= r p r2 Power
   /TITLE "r Effect Size and Power".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= r rFAU rOP
  /TITLE "r and r Bias Adjustments".
```

Table 1. *r*, effect size, and power

Pearson correlation coefficient (*r*)	*p*-value	Variance explained by the relationship (r^2)	Post-hoc power
0.502	0.048	0.252	0.512

Table 2. Estimates for *r* and *r* bias adjustments

Pearson correlation coefficient (*r*)	Fisher approximately unbiased *r* (r_{FAU})	Olkin & Pratt adjusted *r* (r_{OP})
0.502	0.514	0.516

Results

After implementation of the program, the results in Table 1 from this example display the sample-based correlation coefficient (0.502) along with its subsequent *p*-value (0.048); denoting statistical significance at the 0.05 level (note: $p = 0.000$ from the program would default to < 0.001). Additionally, the matrix generated an r^2 effect size (note: applicable when statistical significance is realized) that indicated a substantial amount of the variance, or about 25%, was explained in the bivariate relationship between *X* and *Y*. Also, the model's overall post-hoc power value, which was based on alpha established at 0.05 and the sample size of 16, was expectedly not robust at 0.512, where power ≥ 0.80 is desired in social science research (Nunnally, 1978).

The results from Table 2 exhibit the correlation coefficient and the r_{FAU} and the r_{OP} bias adjustments. As would be expected, the bias-adjusted indices r_{FAU}(0.514) and r_{OP}(0.516) were very comparable, but noticeably higher in value than *r*(0.502) (i.e., > the aforementioned threshold of 0.01 or +0.012 and +0.014, respectively).

Conclusion

Given the information derived from the tables, such as the probable point estimate bias, the program affords users with more accurate estimates, which may provide a study with added robust inferences about the data (i.e., particularly with a small sample size). As noted by Zimmerman et al. (2003) concerning the utility of applying an *r*-based adjustment, "...if one is troubled by the slight bias in the correlation coefficient for normal populations, it is clear that it can be largely eliminated by the Fisher approximate unbiased estimator or by the Olkin and Pratt estimator" (p. 155).

References

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement, 75*(5), 785-804. doi: 10.1177/0013164414557639

Donner, A., & Rosner, B. (1980). On inferences concerning a common correlation coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 29*(1), 69-76. doi: 10.2307/2346412

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507-521. doi: 10.2307/2331838

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3-32.

Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron, 3*, 329-332.

Gorsuch, R. L., & Lehmann, C. S. (2010). Correlation coefficients: Mean bias and confidence interval distortions. *Journal of Methods and Measurement in the Social Sciences, 1*(2), 52-65. doi: 10.2458/v1i2.114

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological), 15*(2), 193-232. Available from http://www.jstor.org/stable/2983768

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the viewpoint of the user. *Journal of the American Statistical Association, 62*(319), 819-841. doi: 10.1080/01621459.1967.10500896

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Olkin, I. (1967). Correlations revisited. In J. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102-128). Chicago, IL: Rand McNally.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics, 29*(1), 201-211. doi: 10.1214/aoms/1177706717

Padilla, M. A., & Veprinsky, A. (2014). Bootstrapped deattenuated correlation: Nonormal distributions. *Educational and Psychological Measurement, 74*(5), 823-830. doi: 10.1177/0013164414531780

Smithson, M. J. (2000). *Statistics with confidence: An introduction for psychologists*. Thousand Oaks, CA: Sage Publications.

Walker, D. A. (2016). *Fisher's approximately unbiased r and Olkin and Pratt's adjusted r for samples < 30* [Computer program].

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica, 24*(1), 133-158. Retrieved from https://www.uv.es/psicologica/articulos1.03/9.ZUMBO.pdf