The Self-Weighting Model Tutorial: Part 1

Abstract – This is the first of a two-parts tutorial on the Self-Weighting Model (SWM), a new weighting model for statistical analysis. SWM allows within/between-set comparisons, producing estimates with a discriminatory power not found through current weighting strategies. The model is applicable to a wide range of statistical problems for which conditional weighted means are required.

Keywords: standard errors, sampling distributions, correlation coefficients, fisher transformations
Published: 03-10-2015; Updated: 10-08-2016

Introduction

This is the first of a two-parts tutorial on the Self-Weighting Model (SWM), a framework for computing weighted means.

The model was first published in *Communications in Statistics - Theory and Methods* (Garcia, 2012a) and presented at SES, New York (Garcia, 2012b). The content of this tutorial is based on these two references.

The Problem

As discussed by Rodgers and Nicewander (1988), Pearson Correlation Coefficient, r, can be expressed in terms of the covariance between two variables (x, y) normalized by their standard deviations

$$r = \frac{cov\ xy}{s_x s_y} \tag{1}$$

(1) shows that any two r's are dissimilar ratios since the $s_x s_y$ product is specific to a sample. These types of ratios are nonadditive. Thus, computing an arithmetic mean from k number of correlations, $\bar{r} = (1/k)\sum_j^k r_j$, is not mathematically valid.

Another evidence of the nonadditivity of Pearson's r can be obtained through a z-score standardization. This is a normalization technique that consists in standardizing a random variable

by subtracting its mean and dividing the result by its standard deviation. The result is a standard score, *z*. Once computed, a distribution of *z* scores can be used to run different tests; for instance, to do a quantile-quantile analysis (Garcia, 2015c).

What is relevant to the present discussion is that for a set of paired *z* scores, Pearson's *r* is a cosine (Rodgers and Nicewander, 1988); i.e., $r = cos(\theta)$. Said cosine can then be taken for a similarity score—the so-called *cosine similarity*, a measure frequently used in information retrieval and data mining.

As Rodgers and Nicewander stated, it is interesting to point out what happens when the *x*, *y* variables are converted into *z* scores: "When we standardize the two raw variables, the standard deviations become unity and the slope of the regression line *becomes* the correlation."

In that case, Pearson's *r* is a slope. To convince yourself, plot z_x vs. z_y or vice versa. The slope of the corresponding regression curve is Pearson's *r*. If one variable is the rank of the other and there are no ties, slope = $r = r_s$, where r_s is Spearman's Correlation Coefficient. Because cosines, similarities, and slopes are not additive, *z*-score standardizations essentially confirm the nonadditivity of correlation coefficients.

Previous Attempts at Averaging Correlations

If we still want to compute an average statistic from *r* values expressed as similarity scores we could do so using the following trick of the trade: compute *r* values from *z*-scores. As these are cosine similarities, transform these into distance metrics. Compute an average distance and transform the result back to a similarity score.

However, said solution might not always be prescribed. As noted by Lin, distance-similarity transformations cannot be applied arbitrarily (Lin, 1998; Garcia, 2015d). Arbitrary transformations of any kind can ignore variability information hidden in data sets and induce to error. In part 2 of this tutorial, we demonstrate this point.

To overcome the problem of averaging correlations, several strategies have been proposed over the years. For instance, back in 2012 we found out that a vendor of a statistical software, StatSoft and currently owned by Dell (2016), suggested converting *r* values into coefficients of determinations, $R_j = r_j^2$, which are additive, or into Fisher *Z* scores, $Z_j = 0.5\ln[(1+r_j)/(1-r_j)]$, which are also additive.

The latter is known as *Fisher's Z Transformation* (Fisher, 1921). Once either approach is adopted, averages of the form $\bar{R} = (1/k)\sum_j^k R_j$ and $\bar{r} = \sqrt{\bar{R}}$, or of the form $\bar{Z} = (1/k)\sum_j^k Z_j$ and $\bar{r} = (e^{\overline{Z}j} - e^{\overline{-Z}j})/(e^{\overline{Z}j} + e^{\overline{-Z}j})$ are computed (Silver & Dunlap, 1987; Zsak, 2006).

Other weighting strategies can be found in the meta-analysis literature, specifically when correlations are taken for effect sizes. Currently, the two dominant meta-analysis models are the Hunter-Schmidt's model (Hunter & Schmidt, 2000; Schmidt, Oh, & Hayes; 2009) and the Hedges-Olkin's fixed effect model (Hedges & Olkin, 1985; Field, 2001).

In Hunter-Schmidt's model, a weighted mean of the form $\bar{r} = \sum_j^k n_j r_j / \sum_j^k n_j$ is computed while in the Hedges-Olkin's model a weighted *Fisher Z* score of the form $\bar{Z} = \sum_j^k (n_j - 3) Z_j / \sum_j^k (n_j - 3)$ is computed and, if needed, transformed back into a correlation score. In both models, n_j is the sample size, which is given in terms of the size of the *x*-*y* dataset. In spite of their success, these two meta-analysis models can induce to error.

For instance if a constant sample size is used during a meta analysis, these models return the arithmetic means of the corresponding *r* and *Z* scores! Since correlations are not additive, the results can be challenged. Moreover, if a constant sample size is used with mixed signs, the problem can be compounded. As noted by Field (2003), arithmetic means from correlations with mixed signs and of same sample size can certainly be misleading.

With regard to *Fisher's Z Transformation*, Zimmerman, Zumbo, & Williams (2003) have shown that arbitrarily applying this transformation to correlations, especially from distributions that violate bivariate normality can lead to spurious results, even with large sample sizes.

According to these authors, "…significance tests of hypotheses about validity and reliability coefficients or differences between them require an assumption of bivariate normality despite large sample sizes. Researchers certainly should be aware of this assumption before using the *r* to *Z* transformation in data analysis. If it is not tenable, estimates of non-zero values of correlation coefficients can be extremely biased, and significance tests can be invalid."

In other words, if *x* and *y* are not normally distributed, *Fisher's Z Transformation* (and models based on it; e.g., Hedges-Olkin's model) can produce invalid results. To overcome all these drawbacks, we proposed back in 2012 the Self-Weighting Model (SWM). In the next section we

provide a general description of the model. A rigorous derivation is given in the second part of this tutorial.

SWM: A General Description

The general idea behind SWM consists in constructing families of weighted means from the constituent components of a function. The procedure is a straightforward one.

First, the statistic to be averaged and its own constituent statistical terms are identified. The constituent terms are then used to compute local and global weights. These weights store variability information and are used to compute families of weighted means.

To illustrate, consider (1). Pearson's *r* can be treated as a function consisting of three constituent statistics. Thus, $m = 3$,

- cov_{xy}
- s_x
- s_y

Therefore, there are at least $2^m - 1 = 7$ ways of defining a local weight, *w*. The number of possible weights that multiply *r* are:

- s_y
- s_x
- $s_x s_y$
- $1/cov_{xy}$
- s_y/cov_{xy}
- s_x/cov_{xy}
- $(s_x s_y)/cov_{xy} = 1/r$

Similarly, there are at least $2^m - 1 = 7$ ways of computing a global weight, *g*, from a set of *r* values and, theoretically, equal number of weighted means that can be constructed from this set. Let assume for now that $w = s_y$.

Multiplying (1) by this weight, squaring the result, taking summations from *j* to *k*, multiplying by a global weight [defined in this example as $g = 1/\sum_j^k (w_j)^2$], and taking the square root of the result leads to a weighted mean that turns out to be equivalent to a root mean square ratio (*rms*),

$$\bar{r} = \left[\frac{\sum_j^k (sy_j)^2 (r_j)^2}{\sum_j^k (sy_j)^2} \right]^{1/2} = \frac{\sqrt{\frac{\sum_j^k (sy_j)^2 (r_j)^2}{k}}}{\sqrt{\frac{\sum_j^k (sy_j)^2}{k}}} \tag{2}$$

Applying a similar procedure to a coefficient of variations of the form $cv_x = s_x / \bar{x}$, where $w = \bar{x}$, it can be demonstrated that

$$\overline{cv_x} = \left[\frac{\sum_j^k (s_{xj})^2}{\sum_j^k (\bar{x}_j)^2} \right]^{1/2} = \frac{\sqrt{\frac{\sum_j^k (s_{xj})^2}{k}}}{\sqrt{\frac{\sum_j^k (\bar{x}_j)^2}{k}}} \tag{3}$$

which also reduces to an *rms* ratio. Notice that (2) and (3) are L_2-norm ratios!

Conclusion

A general overview of the nonadditivity of correlation coefficient has been presented. The Self-Weighting Model has been introduced as an alternative to current meta-analysis models and weighting strategies.

At first glance its derivation looks like an arbitrary heuristic approach. Therefore, a formal derivation of the model is presented in the second part of this tutorial.

In the meantime, please note that (2) and (3) define conditional weighted means, each being members of specific families. In Part 2 of this tutorial, we derive these families and show that they are all part of a large class of families (Garcia, 2015e).

We will also examine the contribution of local and global weights to the computed weighted means. These weights effectively capture variability information present in the data sets, but overlooked by the weighting strategies previously discussed.

References

Dell (2016). What are Basic Statistic? Retrieved from http://www.statsoft.com/textbook/basic-statistics/#Correlationso

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. Psychological Methods 6-2:161–180 Retrieved from http://psych.colorado.edu/~willcutt/pdfs/Field_2001.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. (1921). On the probable error of a coefficient of correlation deduced from a small sample. Metron 1:3–32. Retrieved from https://digital.library.adelaide.edu.au/dspace/bitstream/2440/15169/1/14.pdf

Garcia, E. (2012a). The Self-Weighting Model, Communications in Statistics - Theory and Methods, 41:8,1421-1427. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/03610926.2011.654037

Garcia, E. (2012b). True Lies: Search Marketing Data Errors Uncovered. Retrieved from http://web.archive.org/web/20121101095532/http:/sesconference.com/archive/2012/newyork/agenda-day1.php

Garcia, E. (2015c). A Tutorial on Quantile-Quantile Plots. Retrieved from http://www.minerazzi.com/tutorials/quantile-quantile-tutorial.pdf

Garcia, E. (2015d). A Tutorial on Distance and Similarity. Retrieved from http://www.minerazzi.com/tutorials/distance-similarity-tutorial.pdf

Garcia, E. (2015e): The Self-Weighting Model Tutorial: Part 2 Retrieved from http://www.minerazzi.com/tutorials/self-weighting-model-tutorial-part-2.pdf

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: implications for cumulative knowledge in Psychology. International Journal of Selection and Assessment 8:275–292. Retrieved from http://www.biz.uiowa.edu/faculty/fschmidt/meta-analysis/hunter_schmidt_2000_rev.pdf

Lin, D. (1998). An Information-Theoretic Definition of Similarity. ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning. pp. 296-304. Retrieved from https://www.cs.swarthmore.edu/~richardw/cs65-f08/litreview/phyo.pdf See also http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.1832

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician 42-1: 59–66. Retrieved from http://www.stat.berkeley.edu/~rabbee/correlation.pdf See also http://www.jstor.org/stable/2685263?seq=1#page_scan_tab_contents

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 http://www.biz.uiowa.edu/faculty/fschmidt/meta-analysis/Schmidt_Oh_Hayes_2009.pdf

Silver, N. C. & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's Z Transformation be used? Journal of Applied Psychology 72-1:146–148. Retrieved from http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&uid=1987-14534-001

Zimmerman, D. W., Zumbo, B. D. & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. Psicológica 24:133–158. Retrieved from http://www.uv.es/revispsi/articulos1.03/9.ZUMBO.pdf See also http://www.redalyc.org/articulo.oa?id=16924109

Zsak, M. I. (2006). *Decision support for energy technology investments in built environment*. Master thesis, page 54 and Appendix 5. Norwegian University of Science and Technology. Retrieved from http://www.iot.ntnu.no/users/fleten/students/tidligere_veiledning/Zsak_V06.pdf