Discussion of the talks by Benjamini and Steinberg

Philip Tzvi Reiss
University of Haifa
reiss@stat.haifa.ac.il https://sites.google.com/view/philreiss

On the Foundations of Applied Statistics
Samuel Neaman Institute, April 11, 2024

Outline

A possible heuristic way to look at the G$\times$L correction:

- Random lab model may be written as

$$y_{g\ell i} = \mu_g + a_\ell + b_{g\ell} + \varepsilon_{g\ell i},$$

  where last three terms are mutually independent and normally distributed.

- Thus

$$\mathrm{Var}(\bar{Y}_{g_1} - \bar{Y}_{g_2}) = \underbrace{E\left[\mathrm{Var}(\bar{Y}_{g_1} - \bar{Y}_{g_2})|L\right]}_{\sigma_\varepsilon^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} + \underbrace{\mathrm{Var}\left[E(\bar{Y}_{g_1} - \bar{Y}_{g_2})|L\right]}_{2\sigma^2_{G\times L}}$$

- Denominator of test statistic (two-sample t-statistic) is square root of an estimate of the first term; but should be square root of an estimate of the whole thing.

	Don't reject	Reject
Null true	TN $(1-\alpha)$	FP (type I, α)
Null false	FN (type II, β)	TP $(1-\beta)$

- Where do replicability and the "type I replicability error" of Jaljuli et al. (2023) fit into this scheme?
- Significant < replicable < true?
- Can above 2×2 model be extended to a 2×3 table?
  a 3×2 table?
  a $2 \times 2 \times 2$ array?

On questionable random effect assumptions. . .

One might argue that it would be well to restrict the calculation of an indication of instability "as if so-and-so were random" to those cases where so-and-so was indeed random. The writers believe this position often leads to artificially lowered estimates of instability because of the exclusion of sources of variability that were sampled, though perhaps not very randomly or completely. Consequently, we encourage treating effects as randomly sampled in many circumstances where the randomness is at best dubious.

Mosteller and Tukey (1977), p. 124
— in chapter 7 ("Hunting out the real uncertainty") of "the green book"

In spline smoothing, much more dubious random effect assumptions serve as a "convenient fiction" that usefully recasts smoothing parameter selection as a mixed model problem (Ruppert et al., 2003; Wood, 2011).

Outline

YB

DMS

Generalized generalizability

Synthesis

From reliability to generalizability (and what it has to do with design)

A highly idiosyncratic history:

- Spearman (1910): reliability coefficient
- Formulated as “intraclass correlation coefficient” (Fisher, 1936)
- Classical test theory (Lord and Novick, 1968; Fleiss, 1986): true score model $X = T + e$ with T, e random, ICC defined as $\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$
- As models multiplied, so did ICC’s (Shrout and Fleiss, 1979)
- In response to inadequacy of the true score model, Cronbach et al. (1963) proposed *generalizability theory*, which emphasizes study design (next slide)
- Developments by Patterson and Thompson (1971), Laird and Ware (1982), Pinheiro and Bates (2000) made mixed-effects models routine

Generalization theory in a nutshell:

- Focuses on variance components ("multifacet") models such as

$$y_{pij} = \mu + a_p + b_i + c_{pi} + \varepsilon_{pij}$$

  where a_p is random person effect, b_i is random item effect, c_{pi} is random interaction effect (but some effects can be fixed).
- Generalization coefficient, like ICC, is a ratio of part to whole variance, but defined in terms of the "universe of scores" we wish to generalize to.
- An initial "G study", to estimate variance components (in above case, $\sigma_p^2, \sigma_i^2, \sigma_{pi}^2$), informs design of "D study" upon which decisions are based.

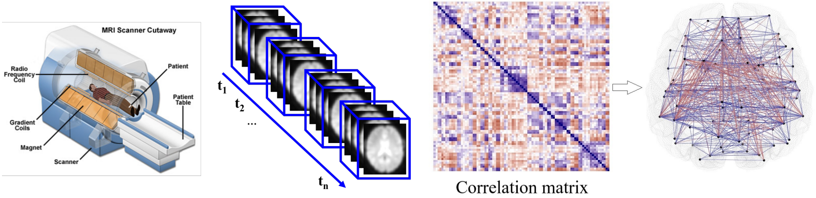
“Experimenters” versus “students of measurement”

From the intro to a classic text on generalizability theory (Cronbach et al., 1972):

> The tester’s neglect of multifacet analysis probably reflects the fact that the design of experiments branched off as a specialty in itself, with the consequence that advances in variance analysis were not brought forcefully to the attention of students of behavioral measurement. The separation was encouraged by the fact that experimenters characteristically regard subjects (persons) as a source of “error” in their analyses, whereas the tester is interested chiefly in the person tested and only secondarily in the conditions of observation. Methodological statements directed to experimenters do not communicate well to students of measurement.

Outline

About fifteen years ago I was involved in early studies of *functional connectivity* in the human brain, as inferred from functional magnetic resonance imaging:



Questions were raised about reliability (and hence reproducibility and generalizability) of findings.

One contribution (Xu et al., 2021):
redefining ICC in terms of distances among observations.

BIOMETRIC METHODOLOGY

Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY
WILEY

Generalized reliability based on distances

Meng Xu[1] | **Philip T. Reiss**[1] | **Ivor Cribben**[2]

[1]Department of Statistics, University of Haifa, Haifa, Israel

[2]Department of Accounting, Operations, and Information Systems, University of Alberta School of Business, Edmonton, Canada

Correspondence
Philip T. Reiss, Department of Statistics, University of Haifa, Haifa 31905, Israel.
Email: reiss@stat.haifa.ac.il

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-2018-06638; Israel Science Foundation, Grant/Award Numbers: 1076/19, 1777/16

Abstract

The intraclass correlation coefficient (ICC) is a classical index of measurement reliability. With the advent of new and complex types of data for which the ICC is not defined, there is a need for new ways to assess reliability. To meet this need, we propose a new distance-based ICC (dbICC), defined in terms of arbitrary distances among observations. We introduce a bias correction to improve the coverage of bootstrap confidence intervals for the dbICC, and demonstrate its efficacy via simulation. We illustrate the proposed method by analyzing the test-retest reliability of brain connectivity matrices derived from a set of repeated functional magnetic resonance imaging scans. The Spearman-Brown formula, which shows how more intensive measurement increases reliability, is extended to encompass the dbICC.

KEYWORDS

functional connectivity, intraclass correlation coefficient, Spearman-Brown formula, test-retest reliability

My former colleagues sought to address the issue via data sharing (`http://fcon_1000.projects.nitrc.org/indi/CoRR/html/`)

Consortium for Reliability and Reproducibility (CoRR) » next | ind

CONSORTIUM FOR RELIABILITY AND REPRODUCIBILITY

Table Of Contents

Quick search

Go

Enter search terms or a module, class or function name.

An open neuroscience

Consortium for Reliability and Reproducibility (CoRR)

The goal of CoRR was to create an open science resource for the imaging community that facilitates the assessment of test-retest reliability and reproducibility for functional and structural connectomics. In order to accomplish this, we have aggregated resting state fMRI (R-fMRI) and diffusion imaging data from laboratories around the world, and are sharing the data via the International Neuroimaging Data-sharing Initiative (INDI). This enables the:

1. Establishment of test-retest reliability and reproducibility for commonly used MR-based connectome metrics
2. Determination of the range of variation in the reliability and reproducibility of these metrics across imaging sites and retest study designs
3. Creation of a standard/benchmark test-retest dataset for the evaluation of novel metrics

Given that this was a retrospective data collection, we have focused on basic phenotypic measures that are relatively standard in the neuroimaging field, as well as fundamental for analyses and sample characterization. Our phenotypic key is organized to reflect three classifications of variables: 1) core (i.e., minimal variables required to characterize any dataset), 2) preferred (i.e., variables that were strongly suggested for inclusion due to their relative import and/or likelihood of being collected by most sites), and 3) optional (variables that are data-set specific or only shared by a few sites).

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Eliminating accidental deviations to minimize generalization error and maximize replicability: Applications in connectomics and genomics

Eric W. Bridgeford[1], Shangsi Wang[1], Zeyi Wang[1], Ting Xu[3], Cameron Craddock[3], Jayanta Dey[1], Gregory Kiar[2], William Gray-Roncal[1], Carlo Colantuoni[1], Christopher Douville[1], Stephanie Noble[4], Carey E. Priebe[1], Brian Caffo[1], Michael Milham[3], Xi-Nian Zuo[5], Consortium for Reliability and Reproducibility, Joshua T. Vogelstein[1,6]*

1 Johns Hopkins University, Baltimore, Maryland, United States of America, **2** McGill University, Montreal, Quebec, Canada, **3** Child Mind Institute, New York, New York, United States of America, **4** Yale University, New Haven, Connecticut, United States of America, **5** State Key Laboratory of Cognitive Neuroscience and Learning, International Data Group/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China, **6** Progressive Learning, Baltimore, Maryland, United States of America

* jovo@jhu.edu

Check for updates

A key paragraph from Bridgeford et al. (2021):

Perhaps the most problematic aspect of these approaches is clear from the popular adage, "garbage in, garbage out" [24]. If the measurements themselves are not sufficiently replicable, then scalar summaries of the data cannot be replicable either. This primacy of measurement is fundamental in statistics, so much so that one of the first modern statistics textbook, R.A. Fisher's, "The Design of Experiments" [25], is focused on taking measurements. Motivated by Fisher's work on experimental design, and Spearman's work on measurement, rather than recommending different post-data acquisition inferential techniques, or computing the repeatability of data after collecting, we take a different approach. Specifically, **we advocate for explicitly and specifically designing experiments to ensure that they provide highly replicable data, rather than hoping that they do and performing post-hoc checks after collecting the data**. Thus, we concretely recommend that new studies leverage existing protocols that have previously been established to generate highly replicable data. If no such protocols are available for your question, we recommend designing new protocols in such a way that replicability is explicitly considered (and not compromised) in each step of the design. Experimental design has a rich history, including in psychology [26] and neuroscience [27, 28]. The vast majority of work in experimental design, however, focuses on designing an experiment to answer a particular scientific question. In this big data age, experiments are often designed to answer many questions, including questions not even considered at the time of data acquisition. How can one even conceivably design experiments to obtain data that is particularly useful for those questions?

Outline

ChatGPT 3.5

You

Hi, I need some bullet points comparing and contrasting the ideas of Yoav Benjamini and David M. Steinberg on replicability and generalizability, can you please help me?

ChatGPT

Here are some bullet points comparing and contrasting the ideas of Yoav Benjamini and David M. Steinberg on replicability and generalizability:

Yoav Benjamini:

- Focuses on controlling the false discovery rate (FDR) to improve replicability.
- Advocates for methods such as the Benjamini-Hochberg procedure to control for multiple comparisons.
- Emphasizes the importance of statistical power and avoiding false positives in scientific research.
- Suggests that replicability can be improved by adjusting p-values to account for multiple testing.

David M. Steinberg:

- Emphasizes the importance of effect size estimation and confidence intervals for generalizability.
- Advocates for reporting effect sizes along with p-values to provide a more comprehensive understanding of the results.
- Highlights the limitations of relying solely on p-values for making inferences about the generalizability of findings.
- Suggests that replicability is enhanced when effect sizes are consistent across studies and can be generalized to different populations or contexts.

In summary, Benjamini's approach focuses on statistical techniques to control for false discoveries and improve replicability, while Steinberg's approach emphasizes effect size estimation and generalizability to enhance the robustness of research findings.

You

Thanks, which works by Steinberg did you use for that?

ChatGPT

The comparison and contrast I provided were based on general principles commonly associated with the ideas of Yoav Benjamini and David M. Steinberg in the context of replicability, generalizability, and effect size estimation. While specific works by David M. Steinberg were not referenced, these concepts are broadly discussed in his contributions to the field of statistics and research methodology. If you're interested in exploring Steinberg's works further, you may consider looking into his papers and books that delve into these topics, such as "Estimation and Inference in Econometrics" and "Statistical Methods for Generalized Linear Models."

You

Do you have full references for those two works, or did you make them up?

ChatGPT

I apologize for any confusion. The references I provided, "Estimation and Inference in Econometrics" and "Statistical Methods for Generalized Linear Models," were hypothetical examples to illustrate the type of works by David M. Steinberg that could be relevant to replicability, generalizability, and effect size estimation. However, I do not have specific information on those exact works or their full references. If you're looking for specific publications by David M. Steinberg, I recommend consulting academic databases such as Google Scholar, PubMed, or university libraries to access his actual works on these topics.

References

Bridgeford, E. W., S. Wang, Z. Wang, T. Xu, C. Craddock, J. Dey, G. Kiar, W. Gray-Roncal, C. Colantuoni, C. Douville, et al. (2021). Eliminating accidental deviations to minimize generalization error and maximize replicability: Applications in connectomics and genomics. *PLoS Computational Biology 17*(9), e1009279.

Cronbach, L., G. Gleser, H. Nanda, and N. Rajaratnam (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Cronbach, L. J., N. Rajaratnam, and G. C. Gleser (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology 16*(2), 137–163.

Fisher, R. A. (1936). *Statistical Methods for Research Workers* (6th ed.). Edinburgh and London: Oliver and Boyd.

Fleiss, J. L. (1986). *Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons.

Jaljuli, I., N. Kafkafi, E. Giladi, I. Golani, I. Gozes, E. J. Chesler, M. A. Bogue, and Y. Benjamini (2023). A multi-lab experimental assessment reveals that replicability can be improved by using empirical estimates of genotype-by-lab interaction. *PLoS Biology 21*(5), e3002082.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Lord, F. M. and M. R. Novick (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.

Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley.

Patterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Shrout, P. E. and J. L. Fleiss (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin 86*(2), 420.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology 3*(3), 271–295.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B 73*(1), 3–36.

Xu, M., P. T. Reiss, and I. Cribben (2021). Generalized reliability based on distances. *Biometrics 77*(1), 258–270.