A mathematical justification for the desirability of the notion of "reproducibility of results"

One might have a vague idea that "reproducibility of results" is a desirable criterion in some branches of science. However, what kind of mathematical justification and support is there for this desirability? The metascience literature does not offer much. If anything in fact, the reproducibility rate and the truth value of a result are not necessarily related [1, 2], which casts doubt whether reproducibility is a necessary component in evaluating the usefulness of results. On the other hand, we believe that a good theoretical understanding of methodological issues related to reproducibility of scientific results must surely be based on a sound mathematical model.

The result we present here is a mathematical justification for the desirability of the notion of "reproducibility of results". It can be defended on solid grounds when all aspects of an idealized study behave mathematically well enough in the sense absence of pathological elements (for example, a model with ill-defined likelihood, a statistical method with non-ideal properties, or nonrandom data).

We start by briefly describing the notion of *idealized study* which is instrumental for our purpose to give a mathematical justification for the concept of reproducibility of results. Details of an early version of the idealized study are given in [2] Appendix 1.

One way to look at an idealized study is as a mathematical structure, in the sense that it is a set endowed with well-specified relationships. Idealized study is useful to investigate the reproducibility of results from a statistical perspective. As common in theoretical statistics work, an idealized study assumes that there is a true probability model generating the data M_T, completely specified by a probability distribution function F_T of random variable X, which is the observable for a phenomenon of interest. Some known background knowledge K partially specifies M_T up to property $\theta \in \Theta$, which denotes unknown and unobservable components –i.e., model parameters– of M_T. A statement that is in principle testable via statistical inference using a simple random and finite sample $\mathbf{X_n} = (X_1, X_2, \cdots, X_n)$, where $X_i \sim F_T$ is made about θ. Candidate mechanisms M_i, inducing distribution functions F_i are formulated. A fixed and known function S is used to extract the information in $\mathbf{X_n}$ pertinent to M_i. S evaluated at $\mathbf{X_n}$ returns $\mathbf{S_n}$, with non-degenerate distribution function $P(\mathbf{S_n} \leq s)$. Formal statistical inference returns a *result* $\{R = d(\mathbf{S_n}, c),\ R \subset \Theta\}$, where c is a user-defined known quantity, and $d(\cdot, \cdot)$ is a fixed and known non-constant decision function which formalizes the statistical inference (by inducing a frequency assessment for a result). The idealized experiment can be succinctly denoted by $\xi := (M, \theta, \mathbf{X_n}, S, K, d)$, and a result of interest obtained from this study will be denoted by R_o.

$\xi^{(i)} := (M^{(i)}, \theta^{(i)}, \mathbf{X_n}^{(i)}, S^{(i)}, K^{(i)}, d^{(i)})$ which differs from ξ only in $K^{(i)}$ and $\mathbf{X_n}^{(i)}$, where points in $\mathbf{X_n}^{(i)}$ are generated independently from F_T, is a replication experiment. By this definition of the replication experiment, a sequence of replication experiments $\xi^{(1)}, \xi^{(2)}, \cdots$ can be taken as exchangeable in their random aspects and conditionally independent of each other, conditional on all aspects but their data $\mathbf{X_n}^{(i)}$.

Using the independence of $\mathbf{X_n}^{(i)}$, [2] established the existence of the true reproducibility rate $\phi \in (0, 1)$ of the result R_o, where ϕ is defined as the proportion of results deemed equivalent to R_o, to the total number of results in an infinite sequence of idealized studies. The best natural estimator for ϕ is

$$\phi_M = M^{(-1)} \sum_{i=1}^{M} \mathbf{I}_{\{R^{(i)} = R_o\}} \tag{1}$$

obtained from a finite sequence $R^{(1)}, R^{(2)}, \cdots, R^{(M)}$, of idealized studies, where $\mathbf{I}_{\{A\}}$ is an indicator function returning 1 if A and 0 otherwise. ϕ_M converges to ϕ as the

number of studies M grows, by the weak law of large numbers.

The convergence built on the estimator 1 is based on the weak law of large numbers and it establishes that with high probability, the estimated reproducibility rate ϕ_M will be close to the true reproducibility rate ϕ. This result is important per se. However, the point convergence provided by the weak law of large numbers does not say quite enough about what we precisely gain, if say, were we to perform more idealized studies. A more reassuring result is convergence as a consequence of the strong law of large numbers. For any ϵ close to zero, and δ close to 1, there exists a M such that

$$P\left(\max_{m\geq M}|\phi_m-\phi|\leq\epsilon\right)\geq\delta. \tag{2}$$

Equation 2 is fundamental for building trust in the concept of reproducibility of results from replication studies. First, it states that if we were to perform replication studies and estimate the reproducibility rate of R_o by ϕ_M from these studies every once in a while, then we are guaranteed that the size of deviations of our estimate from the true reproducibility rate is bounded. That is, equation 2 guarantees that *we are not to be led astray if we continue to perform replication experiments and estimate the reproducibility rate of a result.* Hence, equation 2 gives a theoretical justification of *why we should care about performing more replication studies and be interested in the reproducibility of results.* Second, we also see that appropriate definitions of an idealized study and the reproducibility rate are needed for the result in 2 to hold. An appropriate definition of the reproducibility rate here is taken as the proportion of results of the desired type to the total number of studies. Equation 2 will hold for many other definitions but not for any haphazard definition of the reproducibility rate. For completeness, we consider an example.

The estimator ϕ_M given in equation 1 is based on categorizing –by the indicator $\mathbf{I}_{\{R^{(i)}=R_o\}}$– the result of each idealized study as replicated or not. One might argue that a degree of agreement, as opposed to a 0-1 categorization, between the results of two studies might be a more precise measure of how much a result in a study is reproduced in another. One way to represent this degree of agreement is to replace the indicator function with a function of a continuous random variable. For example, we focus on $Y_{i+1}/Y_i,\ i=1,2,\cdots$ where $Y_i\sim\text{Nor}(0,\sigma)$ is centralized statistic from ξ_i measuring a centralized score on how extreme is a specific result with respect to an original result Y_o. Here Y_i are independent and identically distributed given the background and fixed elements of ξ_i. The statistic $Y_2/Y_1=1$ for example, would tell us that the results in the second and first replication studies have exactly the same degree of agreement. So, we can define the rate of reproducibility as $\phi^*_M=M^{(-1)}\sum_{i=1}^{M}(Y_{i+1}/Y_i)$. Although this definition of a measure of reproducibility looks reasonable, and in fact perhaps preferable to a more crude 0-1 categorization, this is not so. We still have ϕ^*_M converging to ϕ^*, the true reproducibility rate in this context, but for any ϵ^* close to zero, and δ^* close to 1, the existence of an M^* such that

$$P\left(\max_{m\geq M^*}|\phi^*_m-\phi^*|\leq\epsilon^*\right)\geq\delta^* \tag{3}$$

is not guaranteed. Consequently, our argument for the mathematical justification for the concept of reproducibility of results does not hold. This example shows that one has to choose the definition of the rate of reproducibility carefully for reproducibility to be useful measure.

References

1. B. Devezer, L. G. Nardin, B. Baumgaertner, and E. O. Buzbas. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS one*, 14(5):e0216125, 2019.

2. B. Devezer, D. J. Navarro, J. Vandekerckhove, and E. O. Buzbas. The case for formal methodology in scientific reform. *bioRxiv*, 2020.