STATISTICAL REPRODUCIBILITY[1]

Frank P.A. Coolen[2]
Department of Mathematical Sciences, Durham University, UK
E-mail: frank.coolen@durham.ac.uk
Tahani Coolen-Maturi[3]
Department of Mathematical Sciences, Durham University, UK
E-mail: tahani.maturi@durham.ac.uk

1 Statistical Reproducibility

Reproducibility of the results and conclusions of experiments and wider research is essential, and problems with reproducibility have been widely discussed in recent years; an excellent overview of many aspects of reproducibility is provided by Atmanspacher and Maasen [3]. In general, the discussion has mostly been focused on aspects like publication bias and advice on good practice to avoid major reproducibility problems. An issue that has received surprisingly little attention in this discussion is the reproducibility of the results of statistical inference methods, which are often a central part of investigations. The question is straightforward: if an experiment were repeated in the same setting, would it lead to the same conclusion of the statistical analysis as the conclusion based on the data from the original experiment?

Goodman [18] raised the reproducibility issue of statistical inferences, and pointed out a common misunderstanding of the p-value in hypothesis testing, in particular that a small p-value would imply good reproducibility, or replicability as it is called by Goodman. Senn [21] agreed with Goodman that the p-value and reproducibility probability are different measures and that inconsistency between test results from individual studies may be expected. However, he stressed the im-

[1] To appear in: **International Encyclopedia of Statistical Science** (2nd Edition), M. Lovrig (Editor), Springer, Berlin, 2024.

[2] Frank Coolen is Professor of Statistics at the Department of Mathematical Sciences, Durham University (UK). He serves on the editorial boards of, amongst others, *Journal of Statistical Theory and Practice*, *Communications in Statistics*, *Journal of Risk and Reliability*, and *Quality and Reliability Engineering International*. His main research is on foundations and methods of Statistics and Reliability, in particular he has been developing Nonparametric Predictive Inference (NPI) since the mid-90s. Together with collaborators and students, Frank has published over 300 journal and conference papers, a majority of these on NPI or other statistical methods that use imprecise probabilities.

[3] Tahani Coolen-Maturi is Professor of Statistics at the Department of Mathematical Sciences, Durham University (UK). She is Associate Editor for *Journal of Statistical Theory and Practice* and *Mathematical Methods of Statistics*, and was co-editor of two special issues for the former journal. Her main focus is on developing statistical methodologies for real-world applications, such as reliability, finance, and medical applications. Over the past two decades, she has made significant contributions to the development of Nonparametric Predictive Inference (NPI). Tahani has published over 80 journal and conference papers with her collaborators and students, mainly on NPI and related methods to quantify uncertainty via imprecise probability.

portant role the p-value has, and that the p-value and reproducibility probability could be related. In the early years of this century, a number of proposals for a measure of statistical reproducibility were put forward, see Coolen and Bin Himd [10] for a brief overview. While most were ad hoc proposals, an interesting concept was presented by De Martini [17], who uses estimated power as a measure of reproducibility in case a null-hypothesis is rejected. This approach was applied to several basic test scenarios by De Capitani and De Martini [14, 15].

Coolen and Bin Himd [10] presented a fundamentally different approach to quantification of statistical reproducibility. They considered it explicitly as a predictive inference problem, directly in line with the basic question whether or not a hypothetical future experiment, identically performed as the original study, would lead to the same conclusion. This approach is discussed further in the following sections. Billheimer [6] similarly advocated to consider reproducibility as a problem of predictive inference, suggesting a predictive Bayesian approach.

2 Nonparametric Predictive Inference for Statistical Reproducibility

Nonparametric Predictive Inference (NPI) [4, 7, 8, 12] is a frequentist statistics framework based on only few modelling assumptions, with inferences explicitly on future observations, which makes it a particularly suitable methodology for inference on statistical reproducibility. NPI for real-valued observations is based on Hill's assumption $A_{(n)}$ [19] and repeated use of this assumption for inference on multiple future observations. This is a post-data exchangeability assumption, which implies that all orderings of observed data and future data are equally likely. This asssumption is not sufficient to derive precise probabilities for many events of interest, but the maximum lower bound and the minimum upper bound for the probability for an event of interest can be derived by De Finetti's Fundamental Theorem of Probability [16]. These probability bounds are lower and upper probabilities in theory of imprecise probability [5].

The first application of NPI to test reproducibility was presented by Coolen and Bin Himd [10], who presented NPI reproducibility for basic nonparametric tests, such as the Wilcoxon Mann–Whitney test. Senn [21] had reasoned that the reproducibility probability for a hypothesis test may be as low at 0.5 in the worst case, when a test statistic is close to the threshold value between rejection and non-rejection of a null hypothesis. For some basic tests involving a single group of data, or a single population so to say, considered by Coolen and Bin Himd [10], this was confirmed with minimum NPI lower reproducibility probability equal to 0.5. However, for basic tests with two groups of data (two populations), the minimum NPI lower reproducibility probability was less than 0.5, with typically worse reproducibility if the null hypothesis is rejected, with test statistic close to the threshold, than when the null hypothesis is not rejected. This is particularly problematic due to the fact that hypothesis tests tend to be designed in such a way that the real aim of the experiment corresponds to rejection of the null hypothesis. A further worrying fact is that the NPI lower and upper reproducibility probabilities can be relatively small for test statistic values quite far from the threshold.

In recent years, NPI reproducibility has been studied for a range of test scenarios, all confirming the insights mentioned above. These include tests on population quantiles and precendence tests [9], likelihood ratio tests [20], two-sample Kolmogorov-Smirnov test [11], and Student's t-test [22]. Considering the use of t-tests in pharmaceutical product development, Simkus *et al.* [22] also introduced NPI reproducibility of a final decision based on the results of multiple t-tests, which shows that in multiple test scenarios reproducibility can become a major problem.

The main idea of the NPI-based approach to quantify reproducibility of statistical hypothesis tests is to apply the test to the original data, followed by consideration of the test result for all possible future data sets, of the same size as the original data, based on the post-data exchangeability assumption underlying NPI. This leads to computational challenges for all but the most basic tests with small data sets. If, for a given ordering of the future observations among the data observations, without assuming a specific value for an observation in between two original data values, it is possible to determine if the test will lead to certain rejection of the null hypothesis, or certain non-rejection, or that both these conclusions are possible, then sampling of the future orderings provides a good solution which leads to estimates of the NPI lower and upper reproducibility probabilities [13]. If the conclusions of the hypothesis test for a future data set can only be deduced if precise values of the future observations are known, then the NPI bootstrap method can be applied [11, 12].

3 Challenges

The NPI approach to quantification of reproducibility of statistical hypothesis tests has proven to be fruitful and to provide useful insights, serving as warnings with regard to trust in, and interpretation of, test results. Deeper understanding of relations between NPI reproducibility and post-data measures reflecting the strength of statistical inferences, for example power estimates, will be required in order to develop clear guidance for practitioners. A possible approach will be to only go ahead with any process depending on the hypothesis test outcome if NPI reproducibility is sufficiently high. This will require further research on what to do in case of low NPI reproducibility. The NPI approach to reproducibility is not limited to hypothesis testing, and can be explored for other statistical inferences. A first application to reproducibility of estimates of population characteristics is presented in the PhD thesis of Alghamdi [1], where the concept of ϵ-reproducibility is introduced to reflect that a future repeat of an experiment leads to an estimate that differs no more than ϵ from the estimate based on the actual data. An interesting alternative to the nonparametric approach to reproducibility, suitable if one wishes to use a parametric model for the future data observations, is by using a parametric predictive bootstrap method, as presented in the PhD thesis of Aldawsari [2].

The conclusions on test reproducibility, when studied through parametric predictive bootstrap, are largely in line with the NPI reproducibility results. It is also of interest to consider reproducibility quantification using the Bayesian statistics framework, which enables inference for future observations through the posterior predictive distribution.

References

[1] Alghamdi, F.M. (2022). *Reproducibility of Statistical Inference Based on Randomised Response Data.* PhD Thesis, Durham University.

[2] Aldawsari, A. (2023). *Parametric Predictive Bootstrap and Test Reproducibility.* PhD Thesis, Durham University.

[3] Atmanspacher, H. and Maasen, S. (Eds.) (2016). *Reproducibility: Principles, Problems, Practices, and Prospects.* Wiley, Hoboken, New Jersey.

[4] Augustin, T. and Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference* **124**, 251–272.

[5] Augustin, T., Coolen, F.P.A., de Cooman, G. and Troffaes, M.C.M. (Eds.) (2014). *Introduction to Imprecise Probabilities.* Wiley, Chichester.

[6] Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician* **73**, 291-–295.

[7] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters* **36**, 349–357.

[8] Coolen, F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information* **15**, 21–47.

[9] Coolen, F.P.A. and Alqifari, H.N. (2018). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *REVSTAT: Statistical Journal* **16**, 167–185.

[10] Coolen, F.P.A. and Bin Himd, S. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice* **8**, 591-–618.

[11] Coolen, F.P.A. and Bin Himd, S. (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov–Smirnov test. *Journal of Statistical Theory and Practice* **14**, 26.

[12] Coolen, F.P.A. and Coolen-Maturi, T. (2024). Nonparametric predictive inference. In: *International Encyclopedia of Statistical Science*, M. Lovric (Ed.). Springer, Heidelberg.

[13] Coolen, F.P.A. and Marques, F.J. (2020). Nonparametric predictive inference for test reproducibility by sampling future data orderings. *Journal of Statistical Theory and Practice* **14**, 62.

[14] De Capitani, L. and De Martini, D. (2013). Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. *Journal of Statistical Computation and Simulation* **85**, 1056-–1061.

[15] De Capitani, L. and De Martini, D. (2016). Reproducibility probability estimation and RP-testing for some nonparametric tests. *Entropy* **18**, 1—17.

[16] De Finetti, B. (1974). *Theory of Probability* (2 volumes). Wiley, Chichester.

[17] De Martini, D. (2008). Reproducibility probability estimation for testing statistical hypotheses. *Statistics and Probability Letters* **78**, 1056-–1061.

[18] Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine* **11**, 875—879.

[19] Hill, B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association* **63**, 677–691.

[20] Marques, F.J., Coolen, F.P.A. and Coolen-Maturi, T. (2019). Introducing nonparametric predictive inference methods for reproducibility of likelihood ratio tests. *Journal of Statistical Theory and Practice*, **13**, 15.

[21] Senn, S. (2002). A comment on 'a comment on replication, p-values and evidence'. *Statistics in Medicine* **21**, 2437—2444.

[22] Simkus, A., Coolen, F.P.A., Coolen-Maturi, T., Karp, N.A. and Bendtsen, C. (2022). Statistical reproducibility for pairwise t-tests in pharmaceutical research. *Statistical Methods in Medical Research* **31**, 673–688.