Statistical reproducibility for (multiple) pairwise tests in pharmaceutical product development

Andrea Simkus
Durham University Maths Department, UK[a] & AstraZeneca, Cambridge, UK[b]
PhD supervisors: Frank P. A. Coolen[a], Tahani Coolen-Maturi[a], Claus Bendtsen[b] and Natasha A. Karp[b]
andrea.simkus@durham.ac.uk

INTRODUCTION

Statistical reproducibility of scientific tests is a widely discussed topic in pharmaceutical discovery and development, where statistical tests are used to support decision making. This aspect of scientific research has enormous potential. Quantification of reproducibility of statistical tests has the capacity to boost the informative value of empirical experimentation. However, in the existing literature on the topic there has been a lot of confusion about what reproducibility is and how to calculate it. A linked problem is inconsistency and variation in vocabulary employed, as reproducibility, repeatability and replicability, are used interchangeably in related publications.

This posters addresses statistical reproducibility from the perspective of the **nonparametric predictive inference (NPI)**, a frequentist method based on only few assumptions. It is focused on future observations, making it a good approach for inference of reproducibility. **NPI bootstrap** is adopted for calculating the reproducibility. **Statistical reproducibility** aims to answer the question of whether the same decision would be reached if the test was repeated. In this poster a new method is presented for calculating statistical reproducibility for the t**-test**. The method was developed in relation to a test in pharmaceutical discovery and development, which involves 6 test groups whose members are given an increasing dosage of a drug. Multiple pairwise comparisons for the t-test are carried out. First, an algorithm for calculating the reproducibility for the separate pairwise comparisons is presented. Secondly, the question of whether the decision of choosing a particular dose is reproducible is studied.

TEST SCENARIO

The methods presented in this poster were developed in relation to a test scenario in pharmaceutical product development, which investigates how a particular dose of a drug affects a disease. Across 6 groups, members of each group receive a different dose of a treatment against a disease. A chosen variable is observed. Pairwise comparisons on this variable are carried out between groups with doses next to each other. Before reaching a conclusion, p-adjustment is carried out via the Benjamini & Hochberg (BH) procedure (1995). In the test scenario, the null hypothesis is that there is no evidence that the next dose is better than the previous one: $H_0 : \mu = 0$ and the alternative hypothesis is that there is an evidence of the next dose being better than the previous one: $H_1 : \mu > 0$. Here μ stands for population mean. A question is raised whether to reject H_0. There are two possible answers: Yes (Y) or No (N). For the data presented in Table 1, the output is YYYYN.

A	B	C	D	E	F
0.745	0.760	0.220	0.158	-0.101	-0.195
0.858	0.515	0.173	0.264	0.088	-0.052
0.905	0.817	0.270	0.013	-0.052	0.008
0.896	0.784	0.109	0.088	-0.020	-0.042
0.751	0.555	0.190	0.030	-0.018	0.051
1.098	0.631	0.330	0.155	-0.122	-0.004
0.872	0.528	0.415	0.146	-0.010	0.154
0.848	0.555	0.423	0.026	0.139	0.225
	0.627	0.235	0.044	0.194	0.020
	0.689	0.440		-0.044	

Table 1: Data for each dose

STATISTICAL TOOLS

Nonparametric predictive inference (NPI)

Nonparametric predictive inference (NPI) is a frequentist statistical approach, focused on future observations; it employs lower and upper probabilities to quantify uncertainty. It is based on Hill's (post-data) assumption A_n: Assume there are exchangeable random quantities $X_1, ..., X_n$. Their ordered values are $x_{(1)} < x_{(2)} < ... < x_{(n)}$ and let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$ [3]. Here ties have probability 0 [2], i.e. $x_i \neq x_j$ for all $i \neq j$. Then for a future observation X_{n+1}, based on n observations, $A_{(n)}$ is: $P(X_{n+1} \in (x_{(j-1)}, x_{(j)})) = \frac{1}{n+1}$ for $j = 1, 2, ..., n+1$.

NPI-Bootstrap (NPI-B)

Classical NPI takes too much computer time, thus an available method is the NPI-Bootstrap. Before bootstrap algorithm is performed, range is determined. This study uses finite range. To perform NPI-Bootstrap:

1. Create $n + 1$ intervals from n observations.
2. Sample an interval.
3. Continue sampling m further values in the same way to form an NPI-B sample (for the purposes of this paper, $n = m$).
4. Create more NPI-Bs (usually 1000).

APPROACH I: REPRODUCIBILITY OF THE T-TEST FOR SEPARATE PAIRWISE COMPARISONS

The algorithm for calculating NPI bootstrap reproducibility probability (NPI-B-RP) for the t-test has been developed from the adjusted method taken from BinHimd's thesis [1]. The method was originally just used for the Wilcoxon Mann-Whitney (WMW) test.

1. Take the two original samples X and Y, apply the t-test and record the p-value and the decision of this test.
2. From the original data for each dose (X and Y), draw the NPI-B sample and apply the t-test to the bootstrapped samples. Record whether the H_0 is rejected (YES/NO).
3. Repeat Step 2 (each time from original data) 1000 times and then count how many times H_0 is rejected (or not rejected), depending on what was the original decision of this test (see Step 1). This number reports how many times the bootstrap method lead to the original test decision.
4. Divide the output of Step 3 by 1000.
5. Repeat 100 times Steps 2 and 3.
6. Report the mean of the 100 outcomes. This is the NPI-B-RP value.

NPI-B-RP for the WMW test can be calculated by slightly adjusting the algorithm. The algorithm outputs for both tests are presented in Table 2 together with statistics of the original data.

Pairwise comparison	Statistics of the real data: H_0 rejected?	original p-value[a]	effect size	Cohen's d	Algorithm output (Step 6): t-test	WMW test
A vs B	Yes	0.0003	0.226	2.039	0.817	0.798
B vs C	Yes	0.0000	0.366	3.213	0.942	0.926
C vs D	Yes	0.0007	0.178	1.737	0.833	0.824
D vs E	Yes	0.0191	0.097	1.032	0.557	0.585
E vs F	No	0.5977	-0.013	-0.115	0.927	0.940

Table 2: Application of the algorithm for calculating the reproducibility of the t-test and the WMW test

APPROACH II: Reproducibility of the final decision

What matters mostly in deciding what dose of a treatment to go with is the reproducibility for this final decision. In other words, the question is whether the decision of choosing a particular dose is reproducible.

Algorithm

1. Generate 6 new data sets using NPI-B, each one corresponding to one particular dose.
2. For these 6 new data sets adjust the p-value using the BH method.
3. Reach a conclusion; learn which pairwise comparisons are significant and which are not.
4. Obtain a decision output (e.g. YYYYN means do not reject H_0 only for the last pairwise comparison).
5. Repeat steps 1-4 1000 times.
6. Calculate how many times out of 1000 was each decision output reached.

Tree diagram representation

A possible way of representing the outcome is to look at all the different combinations. For the data set, there are 32 possible different combinations, which can be presented in a tree diagram. Figure 1 shows a part of the tree with all the combinations for the adjusted t-test. The reproducibility of the final decision (YYYYN) is 522. That means 522 out of 1000 decisions outputs in Step 4 lead to the original decision YYYYN.

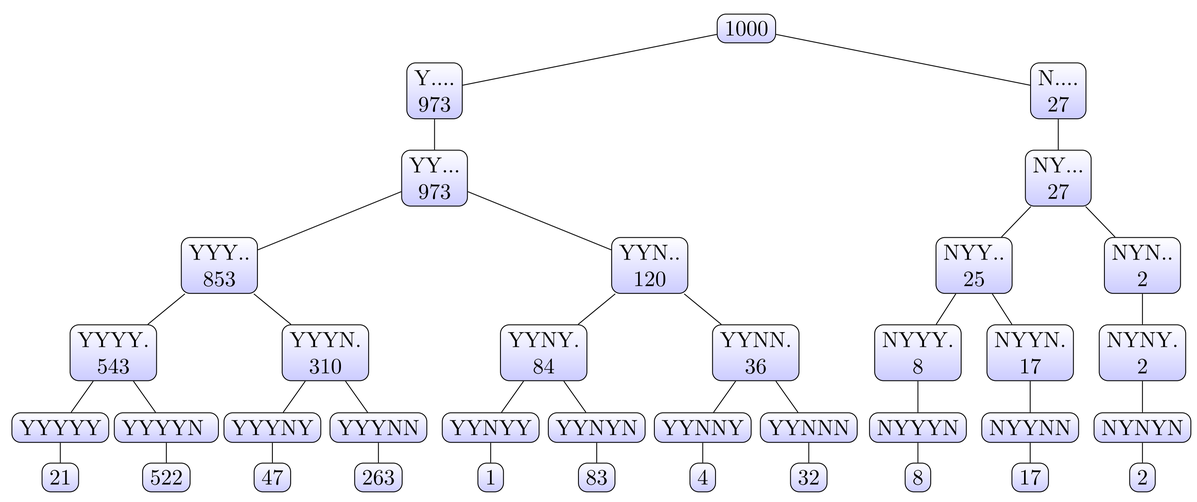

Figure 1: Tree diagram for reproducibility of the final decision for original test scenario (t-test, adjusted p-value)

Imposed decision rules

Previously, we just looked at the occurence of the same results for all the pairwise tests. The **imposed decision rule** allows us to consider more combinations. It defines what combinations to accept. To illustrate it, let's adjust the data for dose D (by adding 1.5 to all the data points before they are logged). Now the original decision made is YYNYN. The imposed decision rule that can be applied to the test scenario with such adjusted data is: **Go from left to right, stop when a null hypothesis is not rejected and take all combinations up to this point.** In Figure 2, this is the branch YYN.. for the decision YYNYN. In this case reproducibility would be 973 out of 1000.

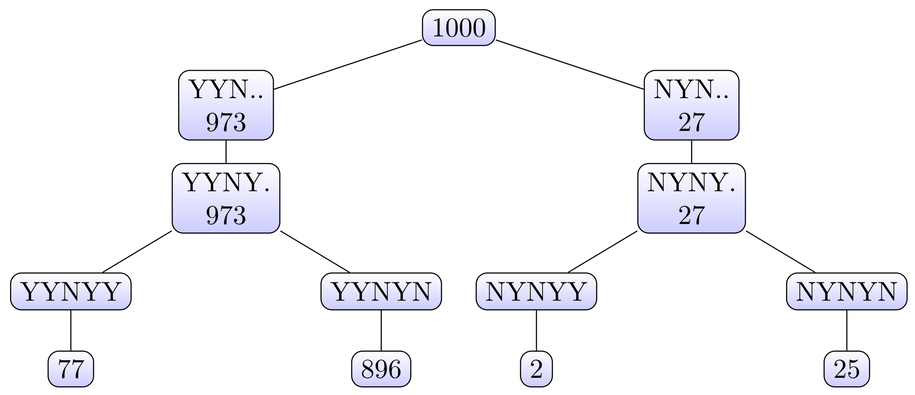

Figure 2: Tree diagram for reproducibility of the final decision for the altered data (t-test, adjsuted p-value)

References

[1] S. BinHimd. *Nonparametric Predictive Methods for Bootstrap and Test Reproducibility*. PhD thesis, Durham University, 2014.

[2] F. P. A. Coolen. On nonparametric predictive inference and objective bayesianism. *Journal of Logic, Language and Information*, 15:21–47, 2006.

[3] F. P. A. Coolen. Nonparametric predictive inference. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 968–970. Springer, 2011.

Acknowledgements

This work was performed under the EPSRC PhD studentship stipend with grant reference number EP/M507854/1. Furthermore, authors gratefully acknowledge support from AstraZeneca for providing data and their context, and for further contribution to the studentship.