Statistical Reproducibility of Hypothesis Tests

Frank Coolen

ICMA-MU Bangkok

18-20 December 2020

Durham University

Reproducibility 'Crisis'

Many reported experimental results are not confirmed when experiments are repeated.

There are many explanations, including publication bias.

The inherent variability of statistical test results has received little attention.

There is very much confusion, for example due to mis-understanding of p-value of a test.

Durham University

Reproducibility of tests

General question:

If a statistical test is repeated under ‘similar’ circumstances, what is the probability that it will lead to the same conclusion?

We consider this a *Prediction* problem, and consider *Frequentist* statistical inference, to base the results as much as possible on the data instead of subjective assumptions.

Focus on ‘conclusion’ being either rejection or non-rejection of a null-hypothesis.

Durham University

Hill's assumption $A_{(n)}$ (Hill, 1968)

- $X_1, \ldots, X_n, X_{n+1}$ are real-valued and exchangeable random quantities
- $x_1 < x_2 < \ldots < x_n$ are the ordered observed values of $X_1, \ldots, X_n$ (and let $x_0 = -\infty$ and $x_{n+1} = \infty$)
- For X_{n+1}, $A_{(n)}$ is given by

$$P(X_{n+1} \in I_j = (x_{j-1}, x_j)) = \frac{1}{n+1} \quad , \quad j = 1, \ldots, n+1$$

Durham University

Nonparametric Predictive Inference (NPI)

- NPI is based on Hill's assumption $A_{(n)}$

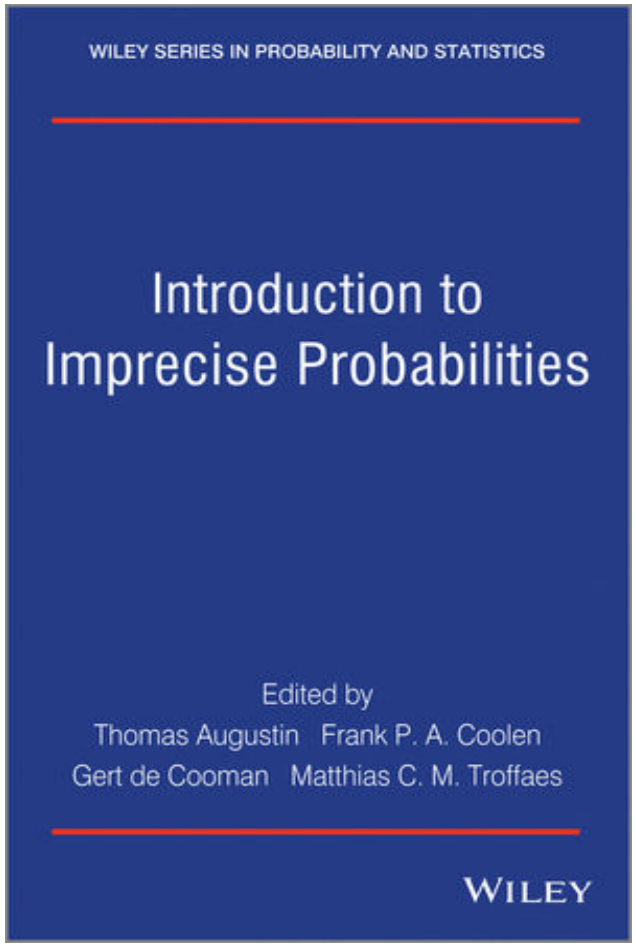
- Let $\mathcal{B}$ be the Borel σ-field over $\mathbb{R}$. For any element $B \in \mathcal{B}$, lower probability $\underline{P}(.)$ and upper probability $\overline{P}(.)$ for the event $X_{n+1} \in B$, based on the intervals $I_j = (x_{j-1}, x_j)$ $(j = 1, 2, \ldots, n+1)$ created by n real-valued non-tied observations, and the assumption $A_{(n)}$, are

$$\underline{P}(X_{n+1} \in B) = \frac{1}{n+1} |\{j : I_j \subseteq B\}|$$

$$\overline{P}(X_{n+1} \in B) = \frac{1}{n+1} |\{j : I_j \cap B \neq \emptyset\}|$$

Durham University

Imprecise Probabilities (2014)



Durham University

Lower and Upper Probabilities

Informal interpretation:

Lower probability $\underline{P}(A)$ reflects the evidence *in favour of* event A

Upper probability $\overline{P}(A)$ reflects the evidence *against* event A

Conjugacy property: $\underline{P}(A^c) = 1 - \overline{P}(A)$

Durham University

NPI for m future observations

- We are interested in $m \geq 1$ future observations, X_{n+i} for $i = 1, \ldots, m$.
- We link the data and future observations via Hill's assumption $A_{(n)}$, actually via $A_{(n+m-1)}$ (which implies $A_{(n+k)}$ for all $k = 0, 1, \ldots, m-2$).
- Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, \ldots, m\}$, then inferences about these m future observations, assuming $A_{(n+m-1)}$, can be based on the following probabilities, for any $(s_1, \ldots, s_{n+1})$ with non-negative integers s_j with $\sum_{j=1}^{n+1} s_j = m$

$$P(\bigcap_{j=1}^{n+1} \{S_j = s_j\}) = \binom{n+m}{n}^{-1}$$

Durham University

Possible interpretation:

Consider a sample of size $n + m$, then randomly select n of them and reveal these first: what can be inferred about the remaining m 'future' values?

All orderings of the $n + m$ values are equally likely, learning the values of the first n does not change this.

Durham University

NPI for Test Reproducibility

Perform hypothesis test on actual data sample of size n, reject H_0 or not.

Prediction of 'future sample' of size n, all orderings of the n future observations among the n actual data observations are equally likely.

For each such ordering of the future observations, consider if H_0 is certainly rejected, possibly rejected and possibly not rejected, or certainly not rejected.

Count the orderings for which the conclusion is certainly the same as for the actual test, this leads to the NPI lower probability for the event that the test result will be reproduced; also including the 'possibly' orderings in the count leads to the corresponding NPI upper probability.

Durham University

NPI-RP for the one-sample signed-rank test

$H_0 : X_1, \ldots, X_n$ symmetrically distributed around median θ.

$$W = \sum_{X_i > \theta} \text{Rank}(|X_i - \theta|)$$

Reject H_0 in favour of H_1 : median $> \theta$ iff $W \geq W_\alpha$, the $100(1-\alpha)$ percentile of the null-distribution for W.

Take $\theta = 0$ (wlog).

NPI considers future observations $X_{n+1}, ..., X_{2n}$. Given real test results $x_{(1)} < ... < x_{(n)}$, there are $\binom{2n}{n}$ equally likely possible orderings of the future observations among the real test results.

Durham University

For each specific ordering, we calculate the minimum and maximum possible test statistic values, $\underline{W}^f$ and $\overline{W}^f$.

If original data led to rejection of H_0, as $W \geq W_\alpha$, then $\underline{RP}$ is the proportion of all $\binom{2n}{n}$ orderings with $\underline{W}^f \geq W_\alpha$ and $\overline{RP}$ the proportion with $\overline{W}^f \geq W_\alpha$.

$\underline{W}^f$ and $\overline{W}^f$ can be calculated without the need to order the n future observations.

For a specific ordering, let S_j be the number of the n future observations in interval $(x_{(j-1)}, x_{(j)})$ (with $x_{(0)} = -\infty, x_{(n+1)} = \infty$).

Durham University

To calculate $\underline{W}^f$, all S_j future observations in $(x_{(j-1)}, x_{(j)})$ are put at ('just to the right of') $x_{(j-1)}$.

Order the absolute data and $-\infty$, with ranks $j = 1, \ldots, n+1$. Let $x_{|j|}$ denote the j-th ordered value if positive, $x_{-|j|}$ if negative ($x_{-|n+1|} = -\infty$).

For $j = 1, \ldots, n+1$, Let T_j be the number of future observations, in the specific ordering considered, that are put at $x_{|j|}$, and T_{-j} the number of such future observations that are put at $x_{-|j|}$. This means that $T_j = S_l$ with $x_{(l-1)} = x_{|j|} > 0$ and $T_{-j} = S_l$ with $x_{(l-1)} = x_{-|j|} < 0$.

$$\underline{W}^f = \sum_{j>0} T_j \left[\frac{(T_j + 1)}{2} + \sum_{|i|<j} T_i \right] \tag{1}$$

$\overline{W}^f$ is similarly derived, with all S_j future observations in $(x_{(j-1)}, x_{(j)})$ put at ('just to the left of') $x_{(j)}$.

Example signed-rank test

sign-ranked data	W	$\underline{RP}$	$\overline{RP}$
1,2,3,4,5,6	21	0.5	1
-1,2,3,4,5,6	20	0.364	0.773
-2,1,3,4,5,6	19	0.326	0.712
-3,1,2,4,5,6	18	0.364	0.718
-2,-1,3,4,5,6	18	0.5	0.788
-4,1,2,3,5,6	17	0.429	0.750
-3,-1,2,4,5,6	17	0.538	0.810
-3,-2,-1,4,5,6	15	0.728	0.902
-6,1,2,3,4,5	15	0.494	0.773
-6,-3,-1,2,4,5	11	0.805	0.935
-6,-5,-4,-3,-2,-1	0	0.992	1

Table: NPI-RP for signed-rank test with H_1 : median>0, $n = 6$, $\alpha = 0.05$, $W_{0.05} = 19$.

Durham University

Tests to compare multiple samples

NPI-reproducibility for tests on two or more samples is done similarly:

Suppose 2 samples, X and Y, with n_x and n_y observations, respectively.

Consider all $\binom{2n_x}{n_x}$ orderings of future X observations among observed X data, and all $\binom{2n_y}{n_y}$ orderings of future Y observations among observed Y data.

Then consider all the $\binom{2n_x}{n_x}\binom{2n_y}{n_y}$ pairs of future X and Y orderings, and investigate the test result for each pair: certainly reject H_0, certainly not reject H_0, or possibly reject H_0. This leads to the lower and upper NPI-RP as before.

Durham University

NPI-RP for basic two-sample precedence test

Precendence testing, typically for lifetime data: one ends the test upon observing the first of a specific ordered observation for the X group or a specific ordered observation for the Y group. There are typically many right-censored observations.

Consider reproducibility for a specific application with test being terminated; no assumptions on the right-censored data at all, lower and upper RP derived as minimum and maximum over all RP values corresponding to the possible orderings of the full data.

Durham University

Example

H_0: X and Y data from same population, tested versus H_1: location shift, i.e. Y values larger. Data: $n_x = 10$ and $n_y = 8$ observations, significance level $\alpha = 0.05$ and testing is terminated on or before the occurrence of the 2nd Y observation.

The test leads to reject H_0 iff $x_{(7)} < y_{(2)}$, and testing is terminated at observation of the first of these.

Durham University

H_0 not rejected

X ranks	Y ranks	$\underline{RP}$	$\overline{RP}$
-	1,2	0.4992	1
1	2,3	0.4951	0.9988
2	1,3	0.4970	0.9992
1,2	3,4	0.4826	0.9924
1,3	2,4	0.4884	0.9946
2,3	1,4	0.4903	0.9951
1-3	4,5	0.4553	0.9733
1-4	5,6	0.4075	0.9314
1-5	6,7	0.3375	0.8582
1-6	7,8	0.25	0.7509
2-7	1,8	0.3663	0.8375

Durham University

H_0 rejected

X ranks	Y ranks	$\underline{RP}$	$\overline{RP}$
1-7	-	0.3833	1
1-6,8	7	0.3367	0.8833
1-5,7,8	6	0.2993	0.8425
1-4,6-8	5	0.2739	0.8098
1-3,5-8	4	0.2593	0.7875
1,2,4-8	3	0.2526	0.7748
1,3-8	2	0.2504	0.7690
2-8	1	0.25	0.7670

Note: The very large imprecision results from the large number of items that are censored, and for which no further assumptions are made.

Durham University

Computational Problem - Solution 1

$\binom{2n}{n}$ orderings of n future observations among n sample observations.

Going through all orderings is only feasible for small n.

For larger n, we can sample orderings and estimate the NPI lower and upper probabilities, using simple random sampling; this also provides confidence intervals.

For each sampled ordering, determine the minimum and maximum possible values for the test statistic, derive estimates for the NPI lower and upper reproducibility probability from these.

Sampling 5,000 or 10,000 orderings is likely to give enough accuracy of the estimates for practical use.

Durham University

Computational Problem - Solution 2

Deriving the minimum and maximum possible values for the test statistic, for a given future ordering, may not be easy.

Instead of the above approach, we can use *NPI Bootstrap* to get an approximate value for the reproducibility.

This does not give lower and upper probabilities, just a single probability.

Durham University

NPI-RP: Ongoing Work with PhD students

Andrea Simkus:
Student t-test; pharmaceutical product development decisions based on multiple tests
(Collaboration with AstraZeneca)

Fatimah Alghamdi:
Reproducibility of hypothesis tests and of estimation with randomised response data

Norah Alalyani:
One-way lay-out tests, including Kruskal-Wallis test and Jonckheere-Terpstra test

Durham University

Main Conclusions

If a test statistic is close to the test threshold, reproducibility is low; it may well be lower than 0.5 for one-sample tests, and possibly substantially lower for two-sample tests.

If a test statistic is not close to the test threshold, reproducibility can still be quite low!

With increased sample size, reproducibility (close to test threshold) does not really get better because future repeat of sample with same size.

Durham University

Main Conclusions

NPI-R is a very different measure compared to estimated power, effect size and other post-data test characteristics.

NPI-RP can be useful to compare different possible tests for the same scenario.

If NPI-RP is low, it is a good idea to repeat the full test, or at least remain aware that any decision resulting from the test must be treated with care.

Durham University

Resources

FC, S. Bin Himd (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests. *JSTP 8*, 591-618.

FC, H.N. Alqifari, (2018). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *REVSTAT - Statistical Journal 16*, 167-185.

F.P.A. Coolen, S. Bin Himd (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov-Smirnov test. *JSTP 14*, article 26.

F.J. Marques, F.P.A. Coolen, T. Coolen-Maturi (2019). Introducing nonparametric predictive inference methods for reproducibility of likelihood ratio tests. *JSTP 13*, article 15.

F.P.A. Coolen, F.J. Marques (2020). Nonparametric predictive inference for test reproducibility by sampling future data orderings. *JSTP 14*, article 62.

Durham University

frank.coolen@durham.ac.uk

npi-statistics.com

Durham University