Reproducibility of some basic nonparametric tests

Frank Coolen

Durham University

with Sulafah Bin-Himd (PhD student)

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?

Bayesian or Frequentist?

Focus on ‘conclusion’ being either rejection or non-rejection of H_0.

Durham University

Some background literature

1992: Goodman: *A comment on replication, p-values and evidence* (Stat. Med.); Discussion by Senn (2002)

2002: Shao & Chow: *Reproducibility probability in clinical trials* (Stat. Med.)

2008: De Martini: *Reproducibility probability estimation for testing statistical hypotheses* (Stat. Prob. Let.)

2012: Begley & Ellis: *Raise standards for preclinical cancer research* (Nature)

and e.g. ...

2002: Posavac: *Using p-values to estimate the probability of a statistically significant replication* (Understanding Statistics)

2005: Killeen: *An alternative to null-hypothesis significance tests* (Psychological Science)

2009: Miller: *What is the probability of replicating a statistically significant effect?* (Psychonomic Bulletin & Review)

Durham University

Sign test

n real-valued iid random quantities $Z_1, \ldots, Z_n$ with median θ, so $P(Z_i < \theta) = P(Z_i > \theta) = 1/2$ for $i = 1, \ldots, n$.

Test $H_0 : \theta = 0$, let $Y = \sum_{i=1}^{n} \mathbf{I}\{Z_i > 0\}$.

Two-sided test with level of significance α: reject H_0 in favour of $H_1 : \theta \neq 0$ iff $Y \geq b_{\alpha/2}$ or $Y \leq n - b_{\alpha/2}$, with $b_{\alpha/2}$ the upper $\alpha/2$ percentile point of Binomial$(n, 1/2)$ distribution.

One-sided test with level of significance α: reject H_0 in favour of $H_1 : \theta > 0$ iff $Y \geq b_\alpha$.

Durham University

NPI-RP for the sign test

Actual test result $Y = y$ positive observations in sample of size n. Now consider a future sample, also of size n, with Y_f denoting the random number of positive observations.

For the one-sided test with $H_1 : \theta > 0$, the relevant NPI lower and upper probabilities, given $Y = y$, are

$$\underline{P}(Y_f \geq b_\alpha | y) =$$
$$1 - \binom{2n}{n}^{-1} \times \left[\binom{2n-y}{n-y} + \sum_{l=1}^{b_\alpha - 1} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right]$$

and

$$\overline{P}(Y_f \geq b_\alpha | y) =$$
$$\binom{2n}{n}^{-1} \times \left[\binom{y+b_\alpha}{y} \binom{2n-y-b_\alpha}{n-y} \right.$$
$$\left. + \sum_{l=b_\alpha+1}^{n} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right]$$

NPI lower and upper reproducibility probabilities:

For $y \geq b_\alpha$, so the original test led to rejection of H_0

$$\underline{RP}(y) = \underline{P}(Y_f \geq b_\alpha | y)$$

and

$$\overline{RP}(y) = \overline{P}(Y_f \geq b_\alpha | y)$$

So future test will also lead to rejection of H_0.

For $y < b_\alpha$, so non-rejection of H_0 in the original test

$$\underline{RP}(y) = \underline{P}(Y_f < b_\alpha | y) = 1 - \overline{P}(Y_f \geq b_\alpha | y)$$

and

$$\overline{RP}(y) = \overline{P}(Y_f < b_\alpha | y) = 1 - \underline{P}(Y_f \geq b_\alpha | y)$$

So future test will also lead to non-rejection of H_0.

Note that we do not consider RP given *only* that H_0 is rejected or accepted; this can be done but taking specific value y into account seems logical.

Example sign test

y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1
1	1.000	1.000
2	1.000	1.000
3	1.000	1.000
4	0.999	1.000
5	0.998	0.999
6	0.995	0.998
7	0.988	0.995
8	0.973	0.988
9	0.947	0.973
10	0.905	0.947
11	0.840	0.905
12	0.750	0.840
13	0.634	0.750
14	0.5	0.634
15	0.5	0.642
16	0.642	0.775
17	0.775	0.882
18	0.882	0.954
19	0.954	0.990
20	0.990	1

Table: Sign test with $H_1 : \theta > 0$, $n = 20$, $\alpha = 0.05$

Durham University

y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1
1	1.000	1.000
2	1.000	1.000
3	1.000	1.000
4	1.000	1.000
5	1.000	1.000
6	0.998	0.999
7	0.995	0.998
8	0.989	0.995
9	0.976	0.989
10	0.952	0.976
11	0.912	0.952
12	0.850	0.912
13	0.760	0.850
14	0.642	0.760
15	0.5	0.642
16	0.5	0.653
17	0.653	0.796
18	0.796	0.909
19	0.909	0.976
20	0.976	1

Table: Sign test with $H_1 : \theta > 0$, $n = 20$, $\alpha = 0.01$

Durham University

y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1
7	0.999	1.000
8	0.998	0.999
9	0.995	0.998
10	0.990	0.995
11	0.981	0.990
12	0.965	0.981
13	0.941	0.965
14	0.904	0.941
15	0.853	0.904
16	0.785	0.853
17	0.702	0.785
18	0.605	0.702
19	0.5	0.605
20	0.5	0.608
21	0.608	0.710
22	0.710	0.801
23	0.801	0.874
24	0.874	0.928
25	0.928	0.964
26	0.964	0.985
27	0.985	0.995
28	0.995	0.999
29	0.999	1.000
30	1.000	1

Table: Sign test with $H_1 : \theta > 0$, $n = 30$, $\alpha = 0.05$

y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1
9	0.999	1.000
10	0.998	0.999
11	0.996	0.998
12	0.991	0.996
13	0.982	0.991
14	0.968	0.982
15	0.945	0.968
16	0.910	0.945
17	0.861	0.910
18	0.794	0.861
19	0.710	0.794
20	0.611	0.710
21	0.5	0.611
22	0.5	0.614
23	0.614	0.724
24	0.724	0.820
25	0.820	0.895
26	0.895	0.948
27	0.948	0.979
28	0.979	0.994
29	0.994	0.999
30	0.999	1

Table: Sign test with $H_1 : \theta > 0$, $n = 30$, $\alpha = 0.01$

Durham University

Two-sided sign test

For the two-sided test with $H_1 : \theta \neq 0$:

if the original test led to rejection of H_0, then

$$\underline{RP}(y) = \underline{P}(Y_f \leq n - b_{\alpha/2} \vee Y_f \geq b_{\alpha/2} | y)$$

et cetera

y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.990	1
1	0.954	0.990
2	0.882	0.954
3	0.775	0.883
4	0.642	0.775
5	0.501	0.644
6	0.495	0.633
7	0.622	0.745
8	0.723	0.827
9	0.787	0.878
10	0.809	0.895
11	0.787	0.878
12	0.723	0.827
13	0.622	0.745
14	0.495	0.633
15	0.501	0.644
16	0.642	0.775
17	0.775	0.883
18	0.882	0.954
19	0.954	0.990
20	0.990	1

Table: Sign test with $H_1 : \theta \neq 0$, $n = 20$, $\alpha = 0.05$

y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.947	1
1	0.829	0.947
2	0.669	0.829
3	0.500	0.669
4	0.500	0.653
5	0.652	0.775
6	0.774	0.863
7	0.862	0.922
8	0.918	0.957
9	0.949	0.976
10	0.959	0.981
11	0.949	0.976
12	0.918	0.957
13	0.862	0.922
14	0.774	0.863
15	0.652	0.775
16	0.500	0.653
17	0.500	0.669
18	0.669	0.829
19	0.829	0.947
20	0.947	1

Table: Sign test with $H_1 : \theta \neq 0$, $n = 20$, $\alpha = 0.01$

y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.998	1
1	0.987	0.998
2	0.960	0.987
3	0.910	0.960
4	0.833	0.910
5	0.734	0.833
6	0.619	0.734
7	0.500	0.620
8	0.500	0.614
9	0.614	0.716
10	0.715	0.800
11	0.800	0.866
12	0.862	0.913
13	0.906	0.944
14	0.932	0.962
15	0.940	0.967
16	0.932	0.962

Table: NPI-RP for sign test with $H_1 : \theta \neq 0$, $n = 30$ and $\alpha = 0.01$.

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 > 0 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 $x_{(0)} = -\infty, x_{(n+1)} = \infty$).

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') $x_{(j)}$.

Durham University

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 $H_1 : \theta > 0$, $n = 6$, $\alpha = 0.05$, $W_{0.05} = 19$.

NPI-RP for the two-sample rank sum test

Same idea: consider all possible future samples which are ordered among the real samples (per sample): then consider all possible pairs of two future samples, and calculate for each such combination the possible values for the test statistic.

Relatively straightforward for one-sided tests, as the 'configurations' that correspond to the NPI lower and upper RP are obvious. Formulae for minimum and maximum value of test statistic for future data in specific ordering (for each sample) have been derived; used to derive $\underline{RP}$ and $\overline{RP}$ as for one-sample signed-rank test.

Two-sided test is more difficult!

Durham University

Further research

Paper with this material to appear in Journal of Statistical Theory and Practice (DOI:10.1080/15598608.2013.819792)

PhD-Thesis by Sulafah Bin-Himd: *Nonparametric predictive methods for bootstrap and test reproducibility*.
(Exam 5 Nov, final thesis online (from my webpage) once approved.)

This includes NPI approach to bootstrapping, which is well suited for NPI-RP in case of larger data sets and for 'less basic' tests.

Interesting further topic: suppose results of several repeated tests are available, or consider multiple future tests (these would not be conditionally independent given outcomes of first test) - RP?