

Reproducibility Probability Estimation for the Wilcoxon-Rank-Sum Test

Stima della probabilità di riproducibilità per il test di Wilcoxon

Daniele De Martini

Dipartimento SEMEQ, Università del Piemonte Orientale

e-mail: demartin@eco.unipmn.it

Riassunto: La stima della probabilità di riproducibilità (RP) di una significatività statistica consiste nella stima a posteriori della potenza, ed è un indicatore dell'affidabilità dei risultati dei test statistici. Si analizzano qui alcuni metodi per la stima, puntuale e per intervalli, della RP per il test di Wilcoxon, e si comparano tramite uno studio di simulazione. Si evince che la soluzione bootstrap fornisce risultati migliori di quella basata sulla normalità asintotica.

Keywords: stima della potenza, intervalli di confidenza per la potenza, risultati di simulazione.

1. Introduction

The Reproducibility Probability (RP) is the power of a statistical test performed on a sample of the same size as the one used in an experiment already performed. It is estimated pointwise by the Observed Power (OP), and De Martini (2005b) argued that OP is a reliability indicator of statistical test results. Moreover, considering as a conservative RP indicator the lower bound of a right-directional confidence interval (see for example Shao and Chow, 2002), a stability criterion to evaluate tests results can be adopted. Hence, RP estimation recently became an important task.

Here, we compare two techniques for estimating the RP of the Wilcoxon-rank-sum test, considering the pointwise and the conservative approach: the Lehmann technique, based on AN approximation of the test statistic, and the bootstrap-plug-in technique, introduced by Collings and Hamilton (1988) and generalized by (De Martini, 2005a).

2. Theoretical recalls

2.1. The power of the Wilcoxon-rank-sum test. Consider two distribution functions $F^{(1)}$ and $F^{(2)}$, and the alternative hypothesis $H_1 : F^{(1)}(t) \geq F^{(2)}(t)$, with the strict inequality for some t (Gibbons and Chakraborti, 2003). Then, without loss of generality, let n be the size of the samples of both groups, and let α be the fixed type I error. Moreover, W_n is the Wilcoxon test statistic, and $w_{n,\alpha}$ is the critical value of the test. Hence, the power of the one-tail Wilcoxon test is:

$$\pi(n, \alpha, F^{(1)}, F^{(2)}) = P_{F^{(1)}, F^{(2)}}(W_n \geq w_{n,\alpha}) = E_{F^{(1)}, F^{(2)}}[\Phi_\alpha^n]$$

where Φ_α^n is the statistical test, i.e. $\Phi_\alpha^n = 1$ if $W_n \geq w_{n,\alpha}$, otherwise $\Phi_\alpha^n = 0$.

2.2. *Available techniques for estimating the power.* In the literature, the problem is not often considered. There is an approximation of the power, provided by Lehmann (1975), based on the asymptotic normal approximation of the test statistic, which has been used by several authors to estimate the power (see for example Collings and Hamilton, 1988). Recently, De Martini (2005a), starting from the plug-in power estimator introduced by Collings and Hamilton (1988), provided a general and consistent method for estimating the power of statistical tests, univariate or multivariate, on one sample or k samples tests, which can be also applied for estimating the power of the Wilcoxon rank-sum test.

2.3. *Lehmann power approximation, and related estimates.* Being $\delta = F^{(1)-1}(0.5) - F^{(2)-1}(0.5)$ the difference between the medians, and σ the common standard deviation, Lehmann asymptotic normal approximation is based on eq.(2.31) in Lehmann (1975):

$$\pi(n, \alpha, F^{(1)}, F^{(2)}) \simeq \phi^{-1} \left(\frac{\delta}{\sigma} \sqrt{\frac{3n^2}{(2n+1)\pi}} - u_{1-\alpha} \right)$$

where ϕ is the cdf of the standard normal, and $u_{1-\alpha} = \phi^{-1}(1 - \alpha)$.

Now, consider given a sample of size n for each group. In order to provide a pointwise estimate applying Lehmann approximation, δ is estimated by the difference between the group averages, i.e. δ_n , and σ by the pooled variance estimator, i.e. σ_n . Then, the Lehmann pointwise estimator is:

$$\phi^{-1} \left(\frac{\delta_n}{\sigma_n} \sqrt{\frac{3n^2}{(2n+1)\pi}} - u_{1-\alpha} \right)$$

In order to provide a lower bound for the power by applying Lehmann approximation, δ/σ (which is the noncentrality parameter of the two sample t -test) is estimated by its lower bound $(\delta/\sigma)_n^\gamma$. Such a bound satisfies $P((\delta/\sigma)_n^\gamma \leq \delta/\sigma) = 1 - \gamma$ and is obtained by solving the following equation:

$$P(T_{2n-2, (\delta/\sigma)_n^\gamma} \leq \sqrt{n}\delta_n/\sigma_n) = 1 - \gamma$$

where $T_{2n-2, (\delta/\sigma)_n^\gamma}$ is a r.v. with a noncentral t distribution with $2n - 2$ dfs and noncentrality parameter $(\delta/\sigma)_n^\gamma$ (see also eq.(18) in Shao and Chow, 2002).

Then, since the power is an increasing function of the noncentrality parameter, the Lehmann lower bound for the power is:

$$\phi^{-1} \left(\left(\frac{\delta}{\sigma} \right)_n^\gamma \sqrt{\frac{3n^2}{(2n+1)\pi}} - u_{1-\alpha} \right)$$

2.4. *Bootstrap-Plug-in power estimation.* Consider given a sample of size n from each group, providing the edfs $F_n^{(1)}$ and $F_n^{(2)}$ respectively. In order to provide a nonparametric pointwise estimator for reproducibility probability, the simple plug-in principle is adopted (see for example Collings and Hamilton, 1988, Efron and Tibshirani, 1993), obtaining

$$\pi(n, \alpha, F_n^{(1)}, F_n^{(2)}) = \frac{1}{n^{2n}} \sum_{i_1=1}^n \cdots \sum_{i_n=1}^n \sum_{j_1=1}^n \cdots \sum_{j_n=1}^n \Phi_\alpha(X_{i_1}^{(1)}, \dots, X_{i_n}^{(1)}; X_{j_1}^{(2)}, \dots, X_{j_n}^{(2)})$$

To provide a lower bound using bootstrap, De Martini (2005b) suggested to consider a sample of size n drawn from $F_n^{(1)}$, of course with replacement, which provides the edf $F_n^{*(1)}$. Analogously, with regard to $F_n^{(2)}$, one obtains $F_n^{*(2)}$. A simple bootstrap p-tile lower bound for the reproducibility probability is obtained by considering the γ th percentile of the distribution of $\pi(n, \alpha, F_n^{*(1)}, F_n^{*(2)})$.

In De Martini (2005a) the consistency of the one sample plug-in power estimator is shown, as well as the asymptotically exact coverage accuracy of the bootstrap lower bound. Extensions to k -sample tests are straightforward.

3. Simulation Study

3.1. Study design. The aim of this work is to compare the performances of these two techniques for estimating RP. We considered $F^{(1)}$ and $F^{(2)}$ with same shape, difference in location δ , and standard deviation $\sigma = 1$. We set a classical sample size: $n = 30$ for each group. Then, we investigated the behavior of estimation techniques with a double sample size $n = 60$. Four different distributions are considered: uniform and normal (which represent classical symmetric distribution shapes), double-exponential (symmetric and heavy tailed), and log-normal (non symmetric). The δ s are set in order to obtain power approximately equal to 90%, 95%. Then, a total of 16 experimental settings are considered: 2 sample sizes $\times$ 4 distributions $\times$ 2 power thresholds.

3.2. Exact power computation. To generate random numbers we used IMSL library, and in particular the following routines: RNUN and ANORIN for uniform, normal, double-exponential and log-normal distributions; RNGDA for discrete distributions.

We performed 10^8 replications of the statistical test in order to approximate the real power with an error lower than 10^{-4} , with a 95% confidence.

To compute the power, we used routine RNKSM of IMSL library, and for continuous distributions we counted the frequency of $\text{stat}(3)$ (i.e. the exact p-values) lower than α , subtracting the frequency of $\text{stat}(3) < \alpha$ and $\text{stat}(1) < \text{stat}(2)$ (in order to compute only right tails).

To compare Lehmann approximation and bootstrap p-tile, we performed 10^4 confidence intervals for everyone of the 16 experimental settings, in order to obtain on the simulated coverage accuracies an error lower than 10^{-2} with a 95% confidence.

3.3. Thresholds for the conservative γ . We computed lower bounds of unidirectional confidence intervals with nominal coverage accuracy of 90%, 95%, according to usual applications. Moreover, we evaluated 75% CIs coverage accuracy, because 75% is a suitable compromise threshold for conservative sample size estimation based on lower bounds of the power (see for example Browne, 1995, Vickers, 2003).

Finally, in order to provide one more pointwise estimate for the bootstrap-plug-in technique, we evaluated its 50% lower bounds.

3.4. Results. On uniform distributed data, the simple bootstrap p-tile works better than Lehmann approximation in every circumstance (see Tab.1), and is almost unbiased. On normal distributed data, Lehmann approximation works better than the simple bootstrap p-tile 10 times over 12, but bootstrap can be improved by studentization (see Efron

and Tibshirani, 1993) or calibration (Hall and Martin, 1988). The difference decreases as n passes from 30 to 60, where the bootstrap works better in 2 circumstances over 6. On double-exponentially distributed data (i.e. with heavy tails), the simple bootstrap p-tile remains almost unbiased, where Lehmann approximation works poorly, underestimating the power. On lognormal distributed data, in experimental settings with power approximately 90%, Lehmann performed very badly where bootstrap performed well as usual. In 23 situations over 24 (4 distributions $\times$ 2 power thresholds $\times$ 3 γ s) bootstrap lower bounds increased their coverage accuracy performances passing from $n = 30$ to $n = 60$. To the contrary, Lehmann technique improved only on uniform and normal distributions.

Table 1: Performances of pointwise and conservative estimators on uniform distributions with $n = 30, 60$ and power approximately equal to 90%.

$n = 30$	$1 - \gamma$	P_w			75%		90%		95%	
		L	PI	B	L	B	L	B	L	B
$\delta =$	CA	.4331	.4860	.4706	.7025	.7250	.8795	.8862	.9385	.9412
0.238	Mean	.8498	.8219	.8273	.7067	.6792	.5413	.5145	.4357	.4125
$\pi =$	Q1	.7821	.7320	.7433	.5389	.4900	.3064	.2600	.1929	.1567
89.96%	Q2	.9257	.9050	.9100	.7732	.7467	.5543	.5233	.4089	.3800
	Q3	.9830	.9770	.9800	.9205	.9100	.7835	.7700	.6593	.6433
$n = 60$	$1 - \gamma$	L	PI	B	L	B	L	B	L	B
$\delta =$	CA	.4532	.4912	.4831	.7236	.7415	.8852	.8922	.9401	.9441
0.165	Mean	.8431	.8220	.8244	.6990	.6739	.5343	.5069	.4306	.4044
$\pi =$	Q1	.7724	.7380	.7400	.5307	.4867	.3011	.2575	.1897	.1533
89.98%	Q2	.9178	.9040	.9067	.7624	.7367	.5443	.5067	.4116	.3633
	Q3	.9796	.9750	.9767	.9135	.9033	.7746	.7533	.6518	.6267

References

- Browne B.J. (1995) On the Use of a Pilot Sample For Sample Size Determination, *Statistics in Medicine*, 14, 1933–1940.
- Collings B.J. Hamilton M.A. (1988) Estimating the power of the Two Sample Wilcoxon Test for Location Shift, *Biometrics*, 44, 847–860.
- De Martini D. (2005a) A general Method to Estimate the Power and the Sample Size of Statistical Tests, *Quaderno Dip.SEMEQ n.80*.
- De Martini D. (2005b) The Emotionality of the p-Value, the Rationality of the Observed Power: How to Measure the Reliability of Statistical Tests Results, *Quaderno Dip.SEMEQ n.82*.
- Efron B., Tibshirani R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- Gibbons J.D., Chakraborti S. (2003) *Nonparametric Statistical Inference*, Dekker, New York.
- Hall P., Martin M.A. (1988) On bootstrap resampling and iteration, *Biometrika*, 75, 661–671.
- Lehmann E.L. (1975) *Statistical Methods Based on Ranks*, Chapman & Hall, New York.
- Shao J., Chow S.C. (2002) Reproducibility Probability in Clinical Trials, *Statistics in Medicine*, 21, 1727–1742.
- Vickers A.J. (2003) Underpowering in Randomized Trials Reporting a Sample Size Calculation, *Journal of Clinical Epidemiology*, 56, 717–720.