Reproducibility Probability Estimation and Testing for some common nonparametric tests

Lucio De Capitani and Daniele De Martini

Abstract The properties of some RP-estimators for some common nonparametric tests are investigated. These estimators can be used to evaluate the variability of the test outcomes and to define the RP-testing decision rule: “accept H_0 if the RP-estimate is lower or equal to 1/2 and reject H_0 otherwise”. We analyze the performances of the considered RP-estimators computing their MSE and evaluating the precision of the aforementioned decision rule, which results very accurate.

Key words: Asymptotic power approximation, Sign Test, Binomial Test, Wilcoxon Signed Rank Test, Kendall Test

1 Introduction

The Reproducibility Probability (RP) is the true power of a statistical test and it can be interpreted as the probability of obtaining a rejection of the null hypothesis in identical subsequent experiments. For this reason, RP-Estimation is an important tool to evaluate the stability and the variability of the test outcomes (see Goodman,1992, and Shao and Chow 2002, for an application to clinical trials). RP-Estimation not only allows to estimate the true power of the test but it can also be used to perform RP-testing. RP-testing has been introduced by De Martini (2008) and it is a technique for testing statistical hypotheses on the basis of the estimate of the RP. The threshold for statistical significance of the RP-based test is $1/2$, with the null hypothesis rejected for high values of RP estimates. In a parametric model, the equivalence between the RP-test and the classical one holds under mild regularity conditions (see De Martini, 2008 for details). In the nonparametric framework,

Lucio De Capitani - Daniele De Martini
Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali, Università degli Studi di Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126, Milano, e-mail: lucio.decapitani1@unimib.it, daniele.demartini@unimib.it

we mainly studied the Wilcoxon Rank-Sum test. In this case, the equivalence between the RP-based test and the classical one still holds in theory, but it cannot be obtained in practice (see De Capitani and De Martini, 2011, for details). Then, we performed a wide simulation study (see De Capitani and De Martini, 2012) to asses the properties of several RP-estimator for the WRS test. We took into account several semi-parametric RP-estimators based on the Asymptotic Normality (AN) of the WRS test statistic and a general nonparametric RP-estimators defined by the plug-in of the empirical distribution functions into the power functional of the WRS test. On one hand, we showed that the probability of disagreement between the RP-based test and the classical one is very low, considering both semi-parametric and nonparametric RP-estimators. On the other hand, we obtained that the general nonparametric RP estimator and the semi-parametric ones have a very similar MSE. These results lead us to a further comparison between semi-parametric and nonparametric RP-Estimators. In detail, in the following we will investigate the performances of the aforementioned RP-estimators for the Sign test, Binomial test, Kendall test, and Wilcoxon Signed Rank Test.

2 RP-Estimation and testing for the Binomial and Sign test

At first we consider the Binomial test. Let $X_1,...,X_n$ be a random sample from the Bernoulli distribution with unknown parameter p. The statistical hypotheses $H_0 : p \leq p_0$ versus $H_1 : p > p_0$ can be tested, with a significance level α, using the exact (and conservative) test $\Psi(X_1,...,X_n) = \begin{cases} 1 & \text{if} \quad \hat{p} > c_\alpha \\ 0 & \text{otherwise} \end{cases}$, where $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $c_\alpha = \frac{b_{(1-\alpha;n,p_0)}}{n}$ and $b_{(q;n,p)}$ is the q-quantile of the binomial distribution with parameters n and p. The test Ψ can be replicated, in analogy to De Martini (2008), using the RP-testing technique. In particular, let $B(\cdot;n,p)$ denote the binomial cdf parameters n and p and let $\hat{p}\bullet$ be the solution of the equation $B(n\hat{p};n,\hat{p}^\bullet) = 1/2$. The RP-estimator $\hat{\pi} = 1 - B(nc_\alpha;n,\hat{p}^\bullet)$ replicates Ψ through the decision rule “accept H_0 if $\hat{\pi} \leq 1/2$ and reject H_0 otherwise”. In place of $\hat{\pi}$, the plug-in RP-estimator $PI = \frac{1}{n^n}\sum_{j_1=1}^{n} \cdots \sum_{j_n=1}^{n} \Psi(X_{j_1},...,X_{j_n})$ can be adopted. With some combinatoric analysis we obtained that PI assumes the values $pi_j = 1 - B\left(nc_\alpha;n,\frac{j}{n}\right)$ with probability $P(PI = pi_j) = \binom{n}{j} p^j (1-p)^{n-j}$. Moreover, it is worthwhile to note that if $\hat{p} = c_\alpha$ the value assumed by PI is $pi_{nc_\alpha} = 1 - B(nc_\alpha;n,c_\alpha) \leq 1/2$. Analogously, the value of PI associated to $\hat{p} = c_\alpha + 1/n$ is $pi_{nc_\alpha+1} = 1 - B(nc_\alpha;n,(nc_\alpha+1)/n) > 1/2$. This demonstrates that the RP-testing rule based on PI and the classical test are equivalent when the Binomial test is applied. Similar theoretical results can be obtained also concerning the well known Sign test, but they are note reported here. We computed the MSE of $\hat{\pi}$ and PI for various sample size, values of p_0, and values of the parameter p. It turns out that, in general, the estimator PI has a lower MSE than $\hat{\pi}$ with a percentage gain of about 5%.

3 RP-Estimation and testing for the Wilcoxon Signed Rank test

Let $X_1,...,X_n$ be a random sample from an arbitrary, continuous, and symmetric cdf F whose median is θ. In order to test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ it is possible to apply the Wilcoxon Signed Rank (WSR) test, which is based on the statistic $W = \sum_{i=1}^{n} I_i R_i = \#\left\{\text{positive}(X_i + X_j)\right\}$ where $R_i = \text{rank}(|X_i|)$ and $I_i = 1$ if $X_i > 0$ and 0 otherwise. The statistic W defines the exact and asymptotic tests

$$\Psi(X_1,...,X_n) = \begin{cases} 1 & \text{if } \ W > w_\alpha \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Psi'(X_1,...,X_n) = \begin{cases} 1 & \text{if } \ W > w'_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where w_α denotes the $(1-\alpha)$-quantile of the exact null distribution of W and $w'_\alpha = \frac{n(n+1)}{4} + z_{1-\alpha}\sqrt{\frac{n(n+1)(2n+1)}{24}}$. As for the WRS test, the exact test Ψ can not be replicated through RP-testing, but it can be well approximated by an RP-based test. In detail we consider the RP-estimators $\hat{\pi} = 1 - \Phi\left[z_{1-\alpha} + \left(\frac{n(n+1)}{4} - W\right)\sqrt{\frac{24}{n(n+1)(2n+1)}}\right]$ and $PI = \frac{1}{n^n}\sum_{j_1=1}^{n} \cdots \sum_{j_n=1}^{n} \Psi'(X_{j_1},...,X_{j_n})$. The estimator $\hat{\pi}$ is based on the AN of W and is obtained applying an approximation procedure similar to those adopted in Noether (1987). It is worthwhile to note that $\hat{\pi}$ replicates the asymptotic WSR test Ψ'. The derivation of the exact distribution of both the RP estimators defined above is not as easy as in the case of the Binomial test. Then, their properties are investigated by a simulation study of 10^4 replications fixing $\alpha = 0.05$ and $n = 15, 30, 60$. For each sample size four different values of θ are considered. The first is $\theta = 0$, the other three values give rise to tests of power approximately equal to 0.2, 0.5 and 0.8. We obtained that the MSE of $\hat{\pi}$ is slightly lower than that of PI with an averaged percentage gain of about 7%. Moreover, we obtained that the test based on $\hat{\pi}$ and the classical exact test Ψ lead to different decisions with an estimated probability of about 0.77%, which is, in fact, quite small. Similarly, the estimated probability of disagreement between Ψ and the RP-testing decision rule based on PI resulted nearly 1.48%.

4 RP-Estimation and testing for the Kendall test of independence

Let (X_i,Y_i), $i = 1,...,n$, be a random sample from a bivariate continuous distribution. To test the independence among X and Y against the alternative of positive association the Kendall test of independence can be adopted. The statistical hypotheses are: $H_0 : \tau = 0$ versus $H_0 : \tau > 0$, where τ is the Kendall rank correlation coefficient. The test statistics is $K = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\text{sign}(X_i - X_j)\cdot\text{sign}(Y_i - Y_j)\right)$ which defines, respectively, the exact and asymptotic tests

$$\Psi(X_1,Y_1...,X_n,Y_n) = \begin{cases} 1 & \text{if } \ K > k_\alpha \\ 0 & \text{otherwise} \end{cases} \text{ and } \Psi'(X_1,Y_1,...,X_n,Y_n) = \begin{cases} 1 & \text{if } \ K > k'_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where k_α denotes the $(1-\alpha)$-quantile of the exact null distribution of W and $w'_\alpha = z_{1-\alpha}\sqrt{\frac{n(n-1)(2n+5)}{18}}$. As for the WSR and WRS tests, the exact Kendall test Ψ cannot be replicated by an RP-based test, but it can be well approximated through RP-testing using the RP-estimators $\hat{\pi} = 1 - \Phi\left(z_{1-\alpha} - K\sqrt{\frac{18}{n(n-1)(2n-5)}}\right)$ and $PI = \frac{1}{n^n}\sum_{j_1=1}^{n}\cdots\sum_{j_n=1}^{n}\Psi'(X_1,Y_1,...,X_n,Y_n)$. Again, $\hat{\pi}$ is based on the AN of K and is obtained by applying an approximation procedure similar to those adopted in Noether (1987). As for the WSR test, $\hat{\pi}$ replicates the asymptotic Kendall test Ψ'. In a simulation study of 10^4 replications we evaluated the properties of the aforementioned estimators. We set $\alpha = 0.05$ and we considered samples of size $n = 15, 30$ and 60 from the bivariate normal distribution with correlation ρ. For every sample size, we choose five different values of ρ. The first is $\rho = 0$, the other four values give rise to tests of power approximately equal to $0.25, 0.4, 0.75$ and 0.9. Simulations highlight that the MSE of $\hat{\pi}$ tends to be larger than that of PI in the scenarios characterized by a low true power of the test. When the true power if high (greater than 0.7) $\hat{\pi}$ works better. Globally the gain of PI with respect to $\hat{\pi}$ is about 1.5%. The exact test Ψ and that based on $\hat{\pi}$ provide different results with a very small estimated probability of 0.46%. Also the estimated probability of disagreement between Ψ and the test based on PI is small, and resulted nearly 0.78%.

5 Conclusions

The analysis we presented in this paper confirm the results obtained in De Capitani and De Martini (2012) for the WRS test. In detail we observed that semi-parametric and non-parametric RP-estimation and testing can be applied also to the Binomial, Sign, WSR, and Kendall tests. The very good performance of the general RP estimator PI must be emphasized. This is a very interesting result since PI is a general RP estimator and it can be applied in almost all the cases of practical interest.

References

1. De Capitani, L., De Martini, D. (2011). On stochastic orderings of the Wilcoxon Rank Sum test statistic - with applications to reproducibility probability estimation testing. *Statistics and Probability Letters*, **81**, 937–946.
2. De Capitani, L., De Martini, D. (2012). Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. Submitted.
3. De Martini, D. (2008). Reproducibility Probability Estimation for Testing Statistical Hypotheses. *Statistics and Probability Letters*, **78**, 1056–1061.
4. Goodman, S.N. (1992). A comment on replication, *p*-values and evidence. *Statistics in Medicine*, **11**, 875–879.
5. Noether, G E. (1987). Sample size determination for some common non-parametric tests. *Journal of the American Statistical Association*, **82**, 645–647.
6. Shao, J., Chow, S.C. (2002). Reproducibility Probability in Clinical Trials. *Statistics in Medicine*, **21**, 1727–1742.