Improving Reproducibility Probability estimation, preserving RP-testing

L. De Capitani, D. De Martini

Department of Statistics and Quantitative Methods, University of Milano-Bicocca

via Bicocca degli Arcimboldi 8, 20126, Milano, Italy.

emails: lucio.decapitani1@unimib.it, daniele.demartini@unimib.it

SUMMARY. Reproducibility Probability (RP) estimation is improved in a general parametric framework, which includes Z, t, χ^2, and F tests. The preservation of RP-testing (i.e. RP estimation based significance testing with threshold at 1/2) is taken into account. Average conservative, weighted conservative, uninformative Bayesian, and Rao-Blackwell RP estimators are introduced, and their relationship studied. Several optimality criteria to define the parameters of weighting functions of conservative RP estimators are adopted. RP-testing holds for average conservative estimators and, under mild conditions, for weighted conservative ones; uninformative Bayesian and Rao-Blackwell RP estimators perform RP-testing only under the location shift model. The performances of RP estimators are compared mainly through MSE. The reduction of MSE given by average conservative estimators is, on average, higher than 20%, and can reach 35%. The performances of optimal weighted RP estimators are even better: on average, the reduction of MSE is higher than 30%.

Keywords: conservative RP estimation; Bayesian RP estimation; average conservative RP estimation; weighted conservative RP estimation.

1 Introduction

Reproducibility is one of the main principles of the scientific method, and several relevant science journals have recently launched a campaign on reproducibility issues, titled "Journals unite for reproducibility" (see [1, 2]). For research findings based on randomness, in particular the many based on statistical test outcomes, the reproducibility probability (RP) should logically be taken into account before considering the reproducibility of methodologies adopted in research. RP estimates are mainly adopted in clinical trials (see e.g. [3, 4, 5]). Nevertheless, RP estimation may be extended to several other experiments where data are analyzed by statistical tests.

Actually, when the type I error probability α is given fixed, RP estimates could replace p-values thanks to RP-testing: this consists in testing statistical hypotheses on the basis of an RP estimate whose significance threshold is $1/2$. Here, we do not argue the adoption of RP estimates to replace p-values: we focus on technical aspects of RP estimation and testing.

Two of the most cited papers on RP estimation are [6] and [3]. RP-testing, introduced several years later [7], holds in many parametric situations. In nonparametrics, RP estimation has been studied for various tests [8, 9, 10]: RP-testing holds exactly for some of them, and with good approximations for the remaining ones.

In practice, RP estimators presented in the above mentioned papers showed the disadvantage of being affected by high variability. For example, the mean error of the simplest RP estimator for the Z-test (i.e. the plug-in pointwise one) is $\simeq .29$ when $RP = .5$, and $\simeq .26$ when $RP = .8$. The mean errors of RP estimators for the other tests studied, parametric or nonpametric, were of the same amplitude.

In this paper, we aim at substantially reducing the MSE of such family of RP estimators. Particular emphasis is given to estimators for which RP-testing holds.

In section 2 some results on the Z-test are presented, with the aim to introduce the basic idea and results of this work. The general framework is defined in section 3 and in section 4 the main results are presented, which are: a class of weighted conservative RP estimators (with some special cases), the uninformative Bayesian RP estimator, a Rao-Blackwell RP estimator, and their theoretical relationship. The behavior of the RP estimators is studied and compared in section 5, where bias, variance and MSE is computed for the t, Z and χ^2 tests, under several settings. An example of application is shown in section 6 and discussion and conclusion follow in section 7.

2 Preliminary results on the Z-test

We started this work focusing on Bayesian RP estimators for the Z-test to compare two means with known variances. We found that the Uninformative Bayesian (UB) estimator showed a variability smaller than that of the simplest pointwise estimator. In Figure 1, the differences between percentiles of RP estimators and RP are reported (i.e. $\widehat{RP}^{\bullet}_{ptile} - RP$), as RP goes from 5% to 95%; the 10th and the 90th percentiles, together with the quartiles (i.e. 25th, 50th, 75th percentiles) of $\widehat{RP}^{\bullet}$ are considered. It is worth noting that the differences between percentiles of $\widehat{RP}^{UB}$ are smaller than those of the pointwise estimator $\widehat{RP}$. Also, $\widehat{RP}^{UB}$ is median biased, whereas $\widehat{RP}$ is unbiased by definition (i.e. $\widehat{RP}_{50\%} - RP = 0, \forall RP$). Nevertheless, we found (numerically) that RP-testing holded for the UB estimator too.

In the light of these findings, we started studying the problem formally, and the results follow here below.

2.1 Basic framework and frequentist RP estimators

Let us consider the test statistic $T_n \sim N(\lambda, 1)$, and assume that the noncentrality parameter λ is equal to 0 under the null hypothesis, whereas $\lambda > 0$ under the alternative one. Being $\alpha \in (0, 1)$ the type I error probability, the critical value is $z_{1-\alpha} = \Phi^{-1}_{0,1}(1 - \alpha)$ and the statistical test is:

$$\Phi_\alpha(T_n) = \begin{cases} 1 & \text{iff} \quad T_n > z_{1-\alpha} \\ 0 & \text{iff} \quad T_n \leq z_{1-\alpha} \end{cases}$$

The power function is $P_\lambda(T_n > z_{1-\alpha}) = \pi_{\alpha,n}(\lambda)$ and, denoting by λ_t the unknown true value of λ, the true power, i.e. the reproducibility probability, is $RP = \pi_{\alpha,n}(\lambda_t)$.

The simplest estimator of the RP is given by plugging the estimator of λ, i.e. T_n, into the power function: $\widehat{RP} = \pi_{\alpha,n}(T_n)$. This pointwise RP estimator is median unbiased, i.e. $P(\widehat{RP} < RP) = 1/2$. Often, lower bounds for the RP are useful, and consequently the γ-conservative estimator for λ was introduced: $\lambda^\gamma_n = T_n - z_\gamma$ [3, 7]. By plugging-in λ^γ_n, the γ-conservative estimator is obtained $\widehat{RP}^\gamma = \pi_{\alpha,n}(\lambda^\gamma_n)$, which gives $P(\widehat{RP}^\gamma < RP) = \gamma$.

2.2 Uninformative Bayesian and Average Conservative RP estimators

When the Bayesian approach is adopted, the posterior distribution of λ_t is first computed, i.e. $h(\cdot|T_n)$. Then, the Bayesian estimator is $\widehat{RP}^B = \int \pi_{\alpha,n}(\lambda) h(\lambda|X_n)\, d\lambda$ (see also [3]). This estimator usually depends on subjective assumptions on the prior distribution of λ_t, and so RP estimate depends on subjectiveness too. For this reason, we consider here only uninformative priors. As a consequence, the posterior distribution of λ_t is normal with mean T_n and unitary variance, that is $h(\cdot|T_n)$ turns out to be equal to the likelihood of λ_t: $\phi_{T_n,1}(\bullet)$. So, the Uninformative Bayesian estimator, which was proposed in [6], results to be

$$\widehat{RP}^{UB} = \int \pi_{\alpha,n}(\lambda)\phi_{T_n,1}(\lambda)\, d\lambda$$

A more intuitive approach to RP estimation, which might be viewed as a robust one, consists in averaging the γ-conservative estimators. In this way, the Average Conservative estimator of the RP turns out to be:

$$\widehat{RP}^{AC} = \int_0^1 \widehat{RP}^{\gamma}\, d\gamma \tag{1}$$

It is worth noting that deviates from T_n in the likelihood function can be written in function of γ, and consequently it is easy to obtain:

$$\widehat{RP}^{UB} = \int \pi_{\alpha,n}(\lambda)\phi_{T_n,1}(\lambda)\, d\lambda = \int_0^1 \widehat{RP}^{\gamma}\, d\gamma = \widehat{RP}^{AC}$$

In other words, in the context of the Z-test the UB estimator can be viewed as the average of frequentist γ-conservative estimators. For clarity, in this section only the notation $\widehat{RP}^{AC}$ will be adopted.

2.3 RP-testing

We formally show here that RP-testing holds in the context of Z-tests, for Average Conservative (and so for Uninformative Bayesian) estimators of the RP.

Theorem 1. *The AC estimator of the RP performs RP-testing.*

$$\Phi_\alpha(T_n) = \begin{cases} 1 & \text{iff} \quad \widehat{RP}^{AC} > 1/2 \\ 0 & \text{iff} \quad \widehat{RP}^{AC} \leq 1/2 \end{cases}$$

Proof. It is a direct consequence of Theorem 1 in [7]. In detail, $\forall\, \gamma \in (0,1)$ we have: $\Phi_\alpha(T_n) = 1$ iff $\widehat{RP}^{\gamma} > \gamma$. Then, by integrating on γ we obtain the claim.

Remark 1 *It is easy to extend Theorem 1 to a wide class of tests, i.e. those included in the general testing framework adopted in [7]. Nevertheless, we do not provide this extention since a more general result will be provided later (i.e. Theorem 2).*

2.4 Variability of the AC/UB RP estimator

The variability of RP estimators is quite high: RP estimates can often fall quite far from RP. The MSEs of $\widehat{RP}^{AC}$ and $\widehat{RP}$ for the Z-test are shown in Figure 1 as functions of RP.

It can be noted that the AC estimator performs better, being its MSE often lower than that of $\widehat{RP}$. On average, the reduction in MSE provided by $\widehat{RP}^{AC}$, with respect to $\widehat{RP}$, is 21.4% (i.e. the relative variation of average MSEs). The best reduction is observed for RP values close to 1/2, where the uncertainty on the decision is highest: the MSE of $\widehat{RP}^{AC}$ is more than 35% smaller than that of $\widehat{RP}$.

The good behavior of $\widehat{RP}^{AC}$ in terms of MSE is given by bias and variance results: in Figure 3 it is shown that the bias of $\widehat{RP}$ (which is median unbiased) is the lowest, and that $\widehat{RP}^{AC}$ presents the lowest variance, but the variance reduction of $\widehat{RP}^{AC}$ is stronger and determines its lower MSE.

Remark 2 *Figures 2 and 3 and the results mentioned above concerning Z-test are general, since the differences $\widehat{RP}^{\bullet}_{ptile} - RP$ of $\widehat{RP}$ and $\widehat{RP}^{AC}$, and consequently their biases, variances and MSEs, do not depend on n or α (proof omitted).*

2.5 Concluding remarks

Given the good results of $\widehat{RP}^{AC}$ in terms of MSE, in the following, the study of AC and UB estimators is extended to a general context, which includes a wide family of tests, in order to: a) generalize UB and AC estimators; b) introduce, if possible, other estimators of the RP; c) study theoretical relationships among different estimators; d) evaluate if RP-testing holds; e) compare the performances of RP estimators, in terms of MSE. To pursue these aims, it is first necessary to provide a general framework.

3 General theoretical framework

Let X be a random variable with distribution function $F_t \in \mathcal{F}$ and let $\mathbf{X}_n$ be a random sample drawn from F_t. Let $T_n = \mathcal{T}(\mathbf{X}_n)$ be the test statistic used to test the statistical hypotheses

$$H_0 : F_t \in \mathcal{F}_0 \qquad \text{vs} \qquad H_1 : F_t \in \mathcal{F} \backslash \mathcal{F}_0. \tag{2}$$

Assume that T_n has a continuous parametric distribution G_{n,λ_t}, with $\lambda_t = \mathcal{L}(F_t, n)$, satisfying the following conditions:

(I) the analytical formula of $G_{n,\lambda}$ is known for all n and $\lambda = \mathcal{L}(n, F)$, $F \in \mathcal{F}$;

(II) for simplicity, and without loss of generality,

$$\sup_{H_0}\{\lambda\} = \sup_{F\in H_0}\{\mathcal{L}(n,F)\} = 0 \qquad \forall\, n;$$

(III) $G_{n,\lambda}$ is stochastically strictly increasing in λ, that is

$$G_{n,\lambda'}(t) > G_{n,\lambda''}(t) \;\; \forall t, \qquad \text{if } \lambda' < \lambda'' \,.$$

Assuming that the statistical hypotheses (2) are one-sided (for example on the left tail), the critical region based on the T_n, thanks to (I)-(III), results

$$\Phi_\alpha\left(\mathbf{X}_n\right) = \begin{cases} 1 & \text{iff } T_n > t_{n,1-\alpha} \\ 0 & \text{iff } T_n \le t_{n,1-\alpha} \end{cases} \quad , \tag{3}$$

where $\alpha \in (0,1)$ is the prefixed Type-I error probability and $t_{n,1-\alpha} = G_{n,0}^{-1}(1-\alpha)$. Note that Z-tests, as well as t-tests, χ^2-tests and F-tests, are included in this framework.

Let Λ be the image of $\mathcal{F}$ under the functional $\mathcal{L}(n,\cdot)$. The power function of test (3) is the function with domain Λ defined as

$$\pi_{\alpha,n}(\lambda) = P_F(T_n > t_{n,1-\alpha}) = 1 - G_{n,\lambda}(t_{n,1-\alpha}) \,. \tag{4}$$

Note that, thanks to (III), $\pi_{\alpha,n}(\lambda)$ is strictly increasing in λ over Λ and the test $\Phi_\alpha\left(\mathbf{X}_n\right)$ is strictly unbiased. The Reproducibility Probability (RP) of the test (3) coincides with the "true power" of the same test:

$$RP = \pi_{n,\alpha}(\lambda_t) = 1 - G_{n,\lambda_t}(t_{n,1-\alpha}) \quad . \tag{5}$$

The "naive" RP estimator, proposed by [3], is obtained through the estimation of λ_t by T_n:

$$\widehat{RP}^N = 1 - G_{n,T_n}(t_{n,1-\alpha}) \,, \tag{6}$$

The γ-conservative estimator of λ_t, denoted by $\hat{\lambda}_n^\gamma$, is implicitly defined as the solution of $G_{n,\hat{\lambda}_n^\gamma}(T_n) = 1-\gamma$. Consequently, the general γ-conservative RP estimator [7] is:

$$\widehat{RP}^\gamma = \pi_{n,\alpha}\left(\hat{\lambda}_n^\gamma\right) = 1 - G_{n,\hat{\lambda}_n^\gamma}(t_{n,1-\alpha}) \,, \tag{7}$$

When the amount of conservativeness adopted is $\gamma = 0.5$, the median unbiased estimator is defined:

$$\widehat{RP} = \widehat{RP}^{0.5} = \pi_{n,\alpha}\left(\hat{\lambda}_n^{0.5}\right) = 1 - G_{n,\hat{\lambda}_n^{0.5}}(t_{n,1-\alpha}) \,, \tag{8}$$

We recall that $\widehat{RP}$ performs RP-testing in the general framework above [7], whereas $\widehat{RP}^N$ does not.

4 RP estimators in the general framework

In this section, several families of RP estimators are introduced, and their theoretical relationships are studied.

4.1 A class of Weighted Conservative RP estimators

The idea of average conservative estimation is developed here, and the possibility of averaging the $\widehat{RP}^{\gamma}$s even not uniformly is introduced with the aim to define, if possible, a more suitable setting of weights. Let us denote by $w(\gamma)$ the weight function, where $w(\gamma) \geq 0$ and $\int_0^1 w(\gamma) d\gamma = 1$. Then, the class of Weighted Conservative estimators is:

$$\widehat{RP}^{WC} = \int_0^1 \widehat{RP}^{\gamma} w(\gamma) d\gamma \tag{9}$$

In general, $\widehat{RP}^{WC}$ does not fulfill RP-testing, but under a simple condition on w it does.

Theorem 2. *If the mean of weights is* $1/2$*, then the WC estimator performs RP-testing.*

$$\Phi_\alpha(T_n) = \begin{cases} 1 & \text{iff} \quad \widehat{RP}^{WC} > 1/2 \\ 0 & \text{iff} \quad \widehat{RP}^{WC} \leq 1/2 \end{cases}$$

Proof. It is analogous to that of Theorem 1, and exploits $\int_0^1 \gamma\, w(\gamma) d\gamma = 1/2$.

4.1.1 Beta-weighted conservative RP estimators

Let us consider a particular class of weights, where w is the Beta density with parameters a and b:

$$w(\gamma; a, b) = \frac{1}{B(a,b)} \gamma^{a-1} (1-\gamma)^{b-1} \qquad a > 0,\ b > 0,\ 0 < \gamma < 1 \ .$$

The variation of a and b allows w to assume a wide range of shapes. In order to allow RP-testing, the weights with average $1/2$ are of particular interest, and are obtained when $a = b$ - note that with this setting the Beta density is symmetric. Thus, the Beta-weighted conservative estimator is defined just when $a = b$:

$$\widehat{RP}^{\beta WC}(a) = \frac{1}{B(a,a)} \int_0^1 \widehat{RP}^{\gamma} \gamma^{a-1} (1-\gamma)^{a-1} d\gamma \qquad a > 0 \tag{10}$$

Theorem 2 gives that RP-testing holds for $\widehat{RP}^{\beta WC}(a)$.

Remark 3 *When* $a = 1$, w *is uniform, i.e.* $w(\gamma; 1, 1) = 1$, $\forall \gamma \in (0, 1)$, *and the Beta-weighted conservative estimator turns out to be equal to the AC one:* $\widehat{RP}^{\beta WC}(1) = \widehat{RP}^{AC}$, *that is defined in (1).*

Remark 4 *When* a *tends to* ∞, w *degenerates to the Dirac measure on* $1/2$, *so that the Beta-weighted conservative estimator is equal to the pointwise one:* $\lim_{a\to\infty} \widehat{RP}^{\beta WC}(a) = \widehat{RP}$.

Remark 5 *When* a *tends to* 0, w *becomes a discrete measure with support* $\{0,1\}$, *each point with probability mass* $1/2$. *In this case, the Beta-weighted conservative estimator degenerates to a constant:* $\lim_{a\to 0} \widehat{RP}^{\beta WC}(a) = 1/2$ *almost surely. This note will be useful mainly when analyzing results.*

4.1.2 Parameter choices in Beta-symmetrically-weighted conservative RP estimators

The special cases described above emphasize that the choice of a plays a crucial role in the behavior of $\widehat{RP}^{\beta WC}(a)$. Therefore, we introduce several optimality criteria to define some choices of a.

First, the parameter a_{mv} giving the Beta-weighted conservative estimator that minimizes the global variability (i.e. the average of the MSE over $RP \in (0, 1)$) is defined:

$$a_{mv} \text{ is such that } \int_0^1 MSE(\widehat{RP}^{\beta WC}(a_{mv}))\, dRP \leq \int_0^1 MSE(\widehat{RP}^{\beta WC}(a))\, dRP\,, \forall a > 0$$

Second, to provide a more stable estimation for high values of the RP (that are useful, for example, to avoid the second confirmative clinical trial (see [3, 4]) the value a_{mvp} that minimizes the proportional global variability of Beta-weighted conservative estimator, is defined:

$$a_{mvp} \text{ is such that}$$

$$\int_0^1 MSE(\widehat{RP}^{\beta WC}(a_{mvp}))RP\, dRP \leq \int_0^1 MSE(\widehat{RP}^{\beta WC}(a))RP\, dRP\,, \forall a > 0$$

Third, a minimax approach is developed: the parameter a_{mm} giving the minimal maximal variability of Beta-weighted conservative estimator is defined:

$$a_{mm} \text{ is such that } \max_{RP\in(0,1)} MSE(\widehat{RP}^{\beta WC}(a_{mm})) \leq \max_{RP\in(0,1)} MSE(\widehat{RP}^{\beta WC}(a))\,, \forall a > 0$$

Thus, the estimators: $\widehat{RP}^{\beta WC}(a_{mv})$, $\widehat{RP}^{\beta WC}(a_{mvp})$, and $\widehat{RP}^{\beta WC}(a_{mm})$ are defined, and will be considered when comparing the behavior of RP estimators.

Remark 6 *If $MSE(\widehat{RP}^{\beta WC}(a))$ is symmetric for all $a > 0$, then $a_{mv} = a_{mvp}$. This note will be useful mainly when analyzing results. Proof is omitted.*

4.2 The Uninformative Bayesian RP estimator

Being $h(\cdot|T_n)$ the posterior distribution of λ_t, the Bayesian estimator of the RP is $\widehat{RP}^{B} = \int_\Lambda \pi_{n,\alpha}(\lambda)h(\lambda|T_n)\,d\lambda$. When an uninformative prior is adopted, the posterior distribution results $h(\lambda|T_n) = g_{n,\lambda}(T_n)/\int_\Lambda g_{n,\lambda}(T_n)d\lambda$, where $g_{n,\lambda}$ denotes the density of $G_{n,\lambda}$. Thus, the uniformative Bayesian RP-Estimator results:

$$\widehat{RP}^{UB} = C^{-1}\int_\Lambda \pi_{n,\alpha}(\lambda)g_{n,\lambda}(T_n)\,d\lambda \tag{11}$$

where $C = \int_\Lambda g_{n,\lambda}(T_n)d\lambda$.

4.2.1 Relations among Uninformative Bayesian and Weighted Conservative RP estimators

The Bayesian RP-Estimator and the conservative (frequentist) RP-Estimators are linked. Let us consider $G_{n,\lambda}(T_n)$, and looking at it as a function of λ denote it by $L_{n,T_n}(\lambda)$. Then, the conservative estimator of λ results $\widehat{\lambda}_n^\gamma = L^{-1}_{n,T_n}(1-\gamma)$ (note that L^{-1}_{n,T_n} is well defined thanks to (III), which implies that L_{n,T_n} is strictly decreasing and so invertible).

Now, consider the uninformative Bayesian estimator in (11) and the change of variable $\lambda = L^{-1}_{n,T_n}(1-\gamma)$, conditioned to T_n, in the related integral.
Since $d\lambda = d\gamma/{-l_{n,T_n}}(L^{-1}_{n,T_n}(1-\gamma))$, where $l_{n,T_n} = L'_{n,T_n}$, and assuming that the image of Λ under the functional $\mathcal{L}(n,\cdot)$ is $(0,1)$, the UB estimator becomes:

$$\begin{aligned}
\widehat{RP}^{UB} &= C^{-1}\int_0^1 [1 - G_{n,L^{-1}_{n,T_n}(1-\gamma)}(t_{n,1-\alpha})]\frac{g_{L^{-1}_{n,T_n}(1-\gamma)}(T_n)}{-l_{n,T_n}(L^{-1}_{n,T_n}(1-\gamma))}d\gamma \\
&= C^{-1}\int_0^1 [1 - G_{n,\hat{\lambda}_n^\gamma}(t_{n,1-\alpha})]\frac{g_{n,\hat{\lambda}_n^\gamma}(T_n)}{-l_{n,T_n}(\hat{\lambda}_n^\gamma)}d\gamma \\
&= C^{-1}\int_0^1 \widehat{RP}^{\gamma}\frac{g_{n,\hat{\lambda}_n^\gamma}(T_n)}{-l_{n,T_n}(\hat{\lambda}_n^\gamma)}d\gamma
\end{aligned} \tag{12}$$

where $C = \int_\Lambda g_{n,\lambda}(T_n)d\lambda = \int_0^1 \frac{g_{n,\hat{\lambda}_n^\gamma}(T_n)}{-l_{n,T_n}(\hat{\lambda}_n^\gamma)}d\gamma$.

This equation recalls (9) and highlights the relationship between the UB estimator and the WC one. First, note that $w^{UB}(\gamma) = g_{n,\hat{\lambda}_n^\gamma}(T_n)/(-l_{n,T_n}(\hat{\lambda}_n^\gamma))C$, with $\gamma \in (0,1)$, is a weight function since, thanks to (III), $-l_{n,T_n}(\cdot)$ is positive because $L_{n,T_n}(\cdot)$ is strictly decreasing. Nevertheless, $w^{UB}(\gamma)$ is a random weight function, because it depends on T_n.

Consequently, the UB estimator can not be a particular case of the WC one, since the random $w^{UB}(\gamma)$ can not be equal to any prefixed $w(\gamma)$.

Remark 7 *In general, the mean weight of $w^{UB}(\gamma)$ is not $1/2$, so that Theorem 2 concerning RP testing can not be applied. Moreover, in many cases (i.e. with several distributions G) $\widehat{RP}^{UB}$ does not perform RP testing (e.g. $G = t$, $G = \chi^2$).*

Remark 8 *When λ is a location parameter for T_n, $\widehat{RP}^{UB}$ performs RP testing. Indeed, in this case we have $G_{n,\lambda}(t) = G_{n,0}(t-\lambda)$, that implies $dG_{n,0}(t-\lambda)/dt = -dG_{n,0}(t-\lambda)/\lambda$, giving $g_{n,\lambda}(t) = -l_{n,t}(\lambda)$. Consequently, $w^{UB}(\gamma) = 1 = w(\gamma;1,1)$, with $\gamma \in (0,1)$, implying $\widehat{RP}^{UB} = \widehat{RP}^{AC}$, and Theorem 2 holds.*

4.3 An improved Rao-Blackwell RP estimator

In the wake of improving RP estimation, the Rao-Blackwell theorem (see, for example, [11]) is applied here, to reduce the MSE of the naive RP estimator (6).

Given that the noncentrality parameter λ_t can be viewed as a function of the parameters of F_t, i.e. $\lambda_t = \tau(\theta_t)$ where θ_t is a vector of length $d \geq 1$, its plug-in estimator $\tilde{\lambda}_n = \tau\left(\hat{\theta}_n\right)$ can be considered, where $\hat{\theta}_n$ is an estimator of θ_t.

In many technical situations (e.g. when $G_{n,\lambda}$ is Gaussian, or t, χ^2, F, ...) we have that: I) $\hat{\theta}_n$ can be easily defined as a set of sufficient statistics for θ_t; II) the test statistic T_n is a function of the $\hat{\theta}_n$, and is an estimator of the noncentrality parameter: $\tilde{\lambda}_n = T_n$. Under these conditions, when the Rao-Blackwell theorem is applied to (6) an RP estimator with lower variance is obtained:

$$\begin{aligned}
\widehat{RP}^{RB} &= E\left[\widehat{RP}^{N} | \hat{\theta}_n\right] \\
&= \int_{\mathbb{R}^d} \left(1 - G_{n,\tau(\mathbf{t})}(t_{n,1-\alpha})\right) \mathbf{f}_{\hat{\theta}_n}(\mathbf{t}) d\mathbf{t} \\
&= \int_{\mathbb{R}} \left(1 - G_{n,t}(t_{n,1-\alpha})\right) g_{n,\tilde{\lambda}_n}(t) dt \qquad (13)
\end{aligned}$$

The Rao-Blackwellization might be also applied to other RP estimators, e.g. to the median unbiased one ($\widehat{RP}$), but this possibility is not developed here.

4.3.1 Relations among naive-Rao-Blackwellized and Weighted Conservative RP estimators

Consider (13) and the change of variable $\tilde{\lambda}_n = T_n$, that give:

$$\widehat{RP}^{RB} = \int_{\mathbb{R}} \left(1 - G_{n,t}(t_{n,1-\alpha})\right) g_{n,T_n}(t) dt$$

$$= \int_0^1 \left(1 - G_{n,\hat{\lambda}_n^\gamma}(t_{n,1-\alpha})\right) \frac{g_{n,T_n}(\hat{\lambda}_n^\gamma)}{-l_{n,T_n}(\hat{\lambda}_n^\gamma)} d\gamma$$

$$= \int_0^1 \widehat{RP}^\gamma \frac{g_{n,T_n}(\hat{\lambda}_n^\gamma)}{-l_{n,T_n}(\hat{\lambda}_n^\gamma)} d\gamma \tag{14}$$

Hence, even the RB estimator recalls the WC one (9). Nevertheless, $w^{RB}(\gamma) = g_{n,T_n}(\hat{\lambda}_n^\gamma)/(-l_{n,T_n}(\hat{\lambda}_n^\gamma))$, with $\gamma \in (0,1)$, is a random weight function, since it depends on T_n. Consequently, the RB estimator can not be included in the WC family.

Remark 9 *Expression (14) of $\widehat{RP}^{RB}$ recalls that of the Uninformative Bayesian in (12), and, actually, the two formulas coincide under the conditions of Remark 7: when λ is a location parameter and $g_{n,0}(t)$ is symmetric, then $\widehat{RP}^{RB} = \widehat{RP}^{UB} = \widehat{RP}^{AC}$ (see Remark 7), and RP-testing holds (Theorem 2).*

5 Evaluating MSE of RP-estimators

The behavior of several estimators is evaluated here, in terms of MSE, for the t, Z, and χ^2 tests, for three values of α: $0.01, 0.05, 0.1$.

First, $\widehat{RP}^{\beta WC}(a)$ was considered, since it is the most general and flexible estimator; the MSE at eight levels of a within $[0, +\infty]$ was computed, with $RP \in (0,1)$, in order to have a first look at the behavior of $\widehat{RP}^{\beta WC}(a)$ and of the values of a that perform well. The values of a taken into account are: $0, 0.3, 0.6, 1, 1.2, 2.4, 4.8, \infty$; indeed, $a = 1$ gives $\widehat{RP}^{\beta WC}(1) = \widehat{RP}^{AC}$ (which can be considered an intermediate approach), $a = \infty$ gives the extreme estimator $\widehat{RP}^{\beta WC}(\infty) = \widehat{RP}$, and $a = 0$ represents the opposite extreme estimator, $\widehat{RP}^{\beta WC}(0)$.

Then, the following seven estimators are considered:

- the classical pointwise estimator: $\widehat{RP} = \widehat{RP}^{0.5}$
- the average conservative estimator: $\widehat{RP}^{AC}$
- the uninformative Bayesian estimator: $\widehat{RP}^{UB}$
- the Rao-Blackwell estimator: $\widehat{RP}^{RB}$
- the minimal variability Beta-weighted conservative estimator: $\widehat{RP}^{\beta WC}(a_{mv})$

- the RP proportional minimal variability Beta-weighted conservative estimator: $\widehat{RP}^{\beta WC}(a_{mvp})$
- the minimal maximal variability Beta-weighted conservative estimator: $\widehat{RP}^{\beta WC}(a_{mm})$

The "naive" RP estimator in (6) is not considered, since it does not perform RP-testing and its MSE is close the that of $\widehat{RP}$.

Two global indexes based on MSE comparison are computed, concerning MSE reduction provided by the generic RP estimator with respect to the classical pointwise one: the mean relative gain $MRG = \int(MSE(\widehat{RP}^{\bullet})/MSE(\widehat{RP})\,dRP - 1$, and the the relative mean gain $RMG = (\int MSE(\widehat{RP}^{\bullet})\,dRP - \int MSE(\widehat{RP})\,dRP\,)/\int MSE(\widehat{RP})\,dRP$.

Bias and variance are not reported, since the kind of impact they have on RP estimators was shown in Section 2. Since the MSE was evaluated for several settings, just a few graphs are reported here; the remaining results are provided in the supplementary material, as well as graphs on bias and variance.

5.1 MSE for the t-test

RP estimation was studied with $\eta = 10, 30, 50, 100$ dfs. For each of the 12 settings so obtained (i.e 4 ηs times 3 αs), performances were evaluated both for $\widehat{RP}^{\beta WC}(a)$ (when a varies) and for the set of (seven) RP estimators listed above.

As it concerns $\widehat{RP}^{\beta WC}(a)$, results report that variations due to different αs are very small. This was expected for large dfs, since t-distribution can be approximated by the Z one, where MSE does not depend on α (see Section 2), but it holds also for small ηs. Even the differences between MSE as η increases are very small: with $\eta = 10$, the paths of MSE curves are very similar to (just a bit less symmetrical, with respect to $RP = 1/2$, than) those of the Z-test.

First, there is not a value a' for which the MSE of $\widehat{RP}^{\beta WC}(a')$ is lower than that of $\widehat{RP}^{\beta WC}(a)$, with $a \geq 0$, on the whole range of RP.

For the estimators with $a \geq 0.3$ (i.e. all but $\widehat{RP}^{\beta WC}(0)$), the MSE is similar when the RP becomes close to 0 or 1 - actually, it tends to zero; on the contrary, when RP goes from 0.15 to 0.85 the MSEs show large differences, often lying between 0.04 and 0.08 (see Figure 4).

This means that RP estimation can be highly variable, in particular when RP is close to 1/2 (i.e. when the variability of the Bernoullian outcome of a statistical test is maximal), where the MSE is about 0.085 for $\widehat{RP}^{\beta WC}(\infty) = \widehat{RP}$ (this is in accordance with the introductive section); nevertheless, the MSE can be reduced remarkably: for example $\widehat{RP}^{\beta WC}(0.3)$ gives an MSE $\simeq 0.031$ when RP is close to 1/2, providing an MSE reduction of about 63%. The estimator $\widehat{RP}^{\beta WC}(0)$ is almost surely equal to 1/2 and, then, it performs very well when RP is very close to 1/2, and very poorly when RP is close to the margins.

As regards comparisons among the seven estimators here considered, variations in MSE due to different αs are, once again, very small, even when $\eta = 10$. Differences between some estimators are relevant when η is small, since when it increases the MSE of $\widehat{RP}^{UB}$, $\widehat{RP}^{AC}$, and $\widehat{RP}^{RB}$ tend to overlap (indeed, these estimators tend to coincide when $\eta \to \infty$, see Remark 9). Moreover, also $\widehat{RP}^{\beta WC}(a_{mv})$ and $\widehat{RP}^{\beta WC}(a_{mvp})$ tend to coincide as dfs increase, since the MSE curves of $\widehat{RP}^{\beta WC}(a)$ tends to be symmetric when $\eta \to \infty$ (see Remarks 6 and 9).

In general, there is no estimator dominating the others in terms of MSE. Nevertheless, the results are interesting and useful, since the three optimized estimators, i.e. $\widehat{RP}^{\beta WC}(a_{\bullet})$, improved RP estimation not only with respect to $\widehat{RP}$, but also to $\widehat{RP}^{UB}$, $\widehat{RP}^{AC}$ and $\widehat{RP}^{RB}$.

The three latter estimators showed an MSE often lower than that of $\widehat{RP}$, especially when RP is close to 1/2. In particular, $\widehat{RP}^{UB}$ and $\widehat{RP}^{RB}$ work better for high RP and poorly when RP is low; on the contrary, the MSE of $\widehat{RP}^{AC}$ is quite stable ($\simeq 0.06$) when RP goes from 0.15 to 0.85.

The optimized estimators $\widehat{RP}^{\beta WC}(a_{mv})$ and $\widehat{RP}^{\beta WC}(a_{mvp})$ are the best performers, among those considered, when RP is close to 1/2; when RP is high (viz. > 0.85), their behavior is the worst - $\widehat{RP}^{\beta WC}(a_{mvp})$ performs a bit better than $\widehat{RP}^{\beta WC}(a_{mv})$.

The minimax estimator $\widehat{RP}^{\beta WC}(a_{mm})$ does not perform bad close to 0 or 1 (just a bit worse than $\widehat{RP}^{AC}$) and presents an MSE quite a bit lower than $\widehat{RP}^{AC}$ (and not far from the two optimized estimators above) when RP is close to 1/2. In detail, when $\eta = 30$, $\alpha = 0.05$, and RP goes from 0.15 to 0.85, the average MSE reduction with respect to $\widehat{RP}^{AC}$ is 10%, and is 36% with respect to $\widehat{RP}$. Hence, we suggest the adoption of $\widehat{RP}^{\beta WC}(a_{mm})$.

The indexes of MSE reduction (viz. RMG and MRG) of the considered RP-estimators with respect to $\widehat{RP}$, are reported in Tables 1 and 2. It can be noted that although the RMG of $\widehat{RP}^{\beta WC}(a_{mm})$ is not best, its MRG is best, and this estimator shows the most uniform improvement. For completeness, the values of a_{mv}, a_{mvp} and a_{mm}, for the considered values of η and α, are reported in Table 3.

Recall that for the t-test, $\widehat{RP}^{UB}$ and $\widehat{RP}^{RB}$ do not perform RP-testing.

5.2 MSE for the Z-test

The behavior of RP estimators of the Z-test can be viewed as that of the t-test when the dfs go to ∞. It has been noted that, with the Z-test, the behavior of the estimators is independent from α and their MSE curves are symmetric (Remark 2). Then, $\widehat{RP}^{\beta WC}(a_{mv})$ and $\widehat{RP}^{\beta WC}(a_{mvp})$, obtained when $a_{mvp} = 0.13$, coincide (see Remark 6). From Section 2.2 and Remark 9 it also follows that $\widehat{RP}^{UB} = \widehat{RP}^{AC} = \widehat{RP}^{RB}$.

As for the t-test, in general there is no estimator dominating the others in terms of MSE. $\widehat{RP}^{\beta WC}(0.13)$ perform best when RP is close to 1/2, but suffer when $RP \leq .2$ or $RP \geq .8$. Nevertheless, the adoption of $\widehat{RP}^{\beta WC}(a_{mm})$ is suggested, since it still performs better than $\widehat{RP}^{UB} = \widehat{RP}^{AC} = \widehat{RP}^{RB}$, in analogy with the t-test. When RP goes from 0.15 to 0.85, the relative mean gain of $\widehat{RP}^{\beta WC}(a_{mm})$ with respect to that of $\widehat{RP}^{AC}$ is 13%, and is 36% with respect to $\widehat{RP}$.

5.3 MSE for the χ^2-test

Four values of dfs for the χ^2 distribution were considered: $\eta = 4, 9, 16, 30$. For each of the 12 settings (i.e 4 ηs times 3 αs), estimation performances were evaluated in analogy with those of the tests above.

As it concerns Beta-weighted estimators, their MSEs are very similar to those of the t-test, that is, there are small differences varying dfs or α; also, their behavior when the parameter a increases is very close to the related ones under t distributions. Therefore, numerical comments of section 5.1 are still valid, and Figure 4(a) can be considered. Once again, there is no value a' for which the MSE of $\widehat{RP}^{\beta WC}(a')$ is best on the whole range of RP.

Some of the seven estimators considered showed different behaviors compared to previous settings. In particular, $\widehat{RP}^{RB}$ performed very poorely, and $\widehat{RP}^{UB}$ is poor when $RP > 0.5$; moreover, recall that both RP estimators do not perform RP-testing.

In general, $\widehat{RP}$, $\widehat{RP}^{AC}$ and $\widehat{RP}^{\beta WC}(a_{mm})$ performed quite similar to their behaviors under t distributions, whereas $\widehat{RP}^{\beta WC}(a_{mv})$ and $\widehat{RP}^{\beta WC}(a_{mvp})$ are still similar just for $RP > 0.2$. On the contrary, when RP is small, the latter two estimators tend to overestimate quite a bit - this is more evident when α increases. Also, $\widehat{RP}^{UB}$ and, mainly, $\widehat{RP}^{RB}$ are seriously affected by overestimation. This is due to the the domain $(0, +\infty)$ of λ_t, which implies that RP estimates are at least equal to α. The performances of RP estimators are quite similar when η vary.

To conclude, there is no estimator dominating the others in terms of MSE. The RMG and MRG are reported in Tables 4 and 5. Although $\widehat{RP}^{\beta WC}(a_{mm})$ provides, once again, a uniform improvement with respect to $\widehat{RP}$ and therefore seems to be preferable, relative gain indexes indicate that $\widehat{RP}^{\beta WC}(a_{mv})$ is the best performer.

The optimal parameters $a_\bullet$ are quite different from those of the t-test; their values, when η and α vary, are reported in Table 6.

6 Numerical example

Two means are compared to evaluate superiority. Given that the superiority margin of scientific interest is 1, the hypotheses of interest are $H1:\ \mu_T - \mu_C > 1$ vs $H0:\ \mu_T - \mu_C \leq 1$, according to [12]. Two random samples of size 16 are drawn, where the known common variance is 2. The standardized difference between sample means is considered for testing. The experimental error α is set at 2.5%, and the critical value is 1.96. The treatment group showed a mean equal to 2.94 and the control mean resulted 0.79. Thus, the Z statistic resulted significant: $z = \sqrt{16/2}(2.94 - 0.79 - 1)/\sqrt{2} = 2.3 > 1.96$. Also, 2.3 seems quite far from 1.96 and the observed significance appears to be a reliable/stable result.

On the contrary, RP estimates result quite low: $\widehat{rp} = \widehat{rp}^N = 63.31\%$, $\widehat{rp}^{AC} = \widehat{rp}^{UB} = \widehat{rp}^{RB} = 59.50\%$, and $\widehat{RP}^{\beta WC}(0.62) = 58.03\%$. From the RP perspective, this does not look like a stable result. Moreover, the most efficient estimator we introduced provided an estimate quite far from the simplest (naive) one.

With the same observed means, if the sample sizes per group were 32, then $z = 3.25$, the RP estimates are higher: $\widehat{rp} = \widehat{rp}^N = 90.19\%$, $\widehat{rp}^{AC} = \widehat{rp}^{UB} = \widehat{rp}^{RB} = 81.97\%$, and $\widehat{RP}^{\beta WC}(0.62) = 77.93\%$. Now, the most efficient estimators provided an estimate even farther from the naive one. As it concerns conservative estimation, this result is 90%-stable (viz. $\widehat{rp}^{90\%} = 50.45\% > 1/2$): this means that not only significance is observed (viz. $\widehat{rp} > 1/2$), but even the variability-accounted-for 90%-conservative estimate of RP is significant (see [4]).

Assume now that the variance is unknown and two samples of 16 data provide an observed test statistic of $t_{30} = 2.427$. Then, the pointwise RP estimate for the t-test is $\widehat{rp} = 64.38\%$ and the average conservative $\widehat{rp}^{AC} = 60.24\%$, both respecting RP-testing; the uninformative Bayesian and the Rao-Blackwell (that do not fulfill RP-testing) give $\widehat{rp}^{UB} = 61.28\%$ and $\widehat{rp}^{RB} = 61.74\%$; finally, the Beta-weighted estimators with optimized parameters for 30 dfs provide $\widehat{RP}^{\beta WC}(0.11) = 52.94\%$, $\widehat{RP}^{\beta WC}(0.15) = 53.73\%$ and $\widehat{RP}^{\beta WC}(0.59) = 58.47\%$. Once again, the RP values provided by the most efficient estimators are quite far from the simplest one.

7 Discussion and conclusion

We fulfilled the aim of improving RP estimation preserving RP-testing, and our results hold under a quite general model including Z, t, χ^2 and F tests. Several RP estimators have been introduced, some of which stemmed from classical statistical theory, and some other based on original ideas; moreover, their relationship has been studied. In particular, the average conservative Beta-weighted optimized RP estimators (viz. $\widehat{RP}^{\beta WC}(a_{mv})$ and $\widehat{RP}^{\beta WC}(a_{mm})$) has been introduced, for which RP-testing holds and that provided very good numerical results: on average, the MSE reduction provided by them with respect to $\widehat{RP}$ is about 30%.

Since optimal settings of parametrized RP estimators $\widehat{RP}^{\beta WC}(a)$ exists (see Figure 4(a)) but are unknown (depending on the unknown RP), one could resort to *Calibration* (see [13]) to first estimate the best value of a and then use it to estimate the RP. Readers should be informed that calibration work bad in this context, providing RP estimators with higher MSE.

The behavior of RP estimators for local alternatives might be considered in further studies, but it is implicitly already done in this paper. Indeed, the MSE of estimators is computed by keeping fixed the RP, and this condition can be obtained when the noncentrality parameter tends to zero and the sample size increases, i.e. under local alternatives.

We recall that the RP estimators here introduced improve pointwise estimation, whereas in order to evaluate the stability of test outcomes, conservative RP estimation should be adopted (De Martini, 2012).

Further developments might concern extending the average conservative RP estimation techniques here introduced to some nonparametric tests, since several results on nonparametric RP estimation are available (see [9, 10]).

References

[1] Anonymous. Journals unite for reproducibility. *Nature* 2014; **515**(7525): 7–7.

[2] McNutt M. Journals unite for reproducibility. *Science* 2014; **346**(6210): 679–679.

[3] Shao J, Chow SC. Reproducibility Probability in Clinical Trials. *Statistics in Medicine* 2002; **21**(12): 1727–1742.

[4] De Martini D. Stability criteria for the outcomes of statistical tests to assess drug effectiveness with a single study. *Pharmaceutical Statistics* 2012; **11**(4): 273–279.

[5] Hsieh TC, Chow SC, Yang LY. The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics. *Statistics in Medicine* 2013; **32**(3): 406–414.

[6] Goodman SN. A comment on replication, p-values and evidence. *Statistics in Medicine* 1992; **11**(7): 875–879.

[7] De Martini D. Reproducibility Probability Estimation for Testing Statistical Hypotheses. *Statistics and Probability Letters* 2008; **78**(9): 1056–1061.

[8] De Capitani L, De Martini D. On stochastic orderings of the Wilcoxon Rank Sum test statistic-With applications to reproducibility probability estimation testing. *Statistics and Probability Letters* 2011; **81**(8): 937–946.

[9] De Capitani L, De Martini D. Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. *Journal of Statistical Computation and Simulation* 2015a; **85**(3): 468–493.

[10] De Capitani L, De Martini D. Reproducibility probability estimation for common nonparametric testing. *Joint meeting of the International Biometric Society (IBS) Austro-Swiss and Italian Regions* 2015b.

[11] Lehmann EL, Casella G. *Theory of Point Estimation*, Second Edition, Springer: New York, 1998.

[12] Hauschke D, Shall R, Luus HG. Statistical Significance. In: Chow SC editor. Encyclopedia of Biopharmaceutical Statistics. 3rd ed. CRC Press 2012; 1290–1293.

[13] Hall P, Martin MA. On bootstrap resampling and iteration. *Biometrika* 1988; **75**(4): 661–671.

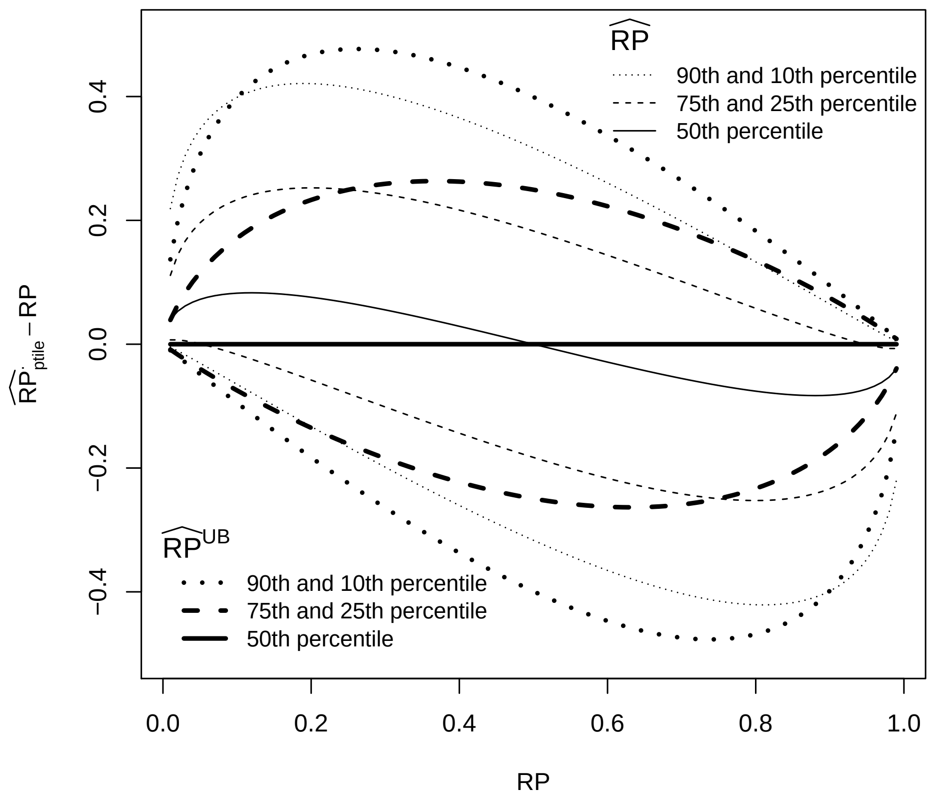


Figure 1: *Differences between percentiles of* $\widehat{RP}^{\bullet}$ *and RP for the Z-test, in function of the RP.*

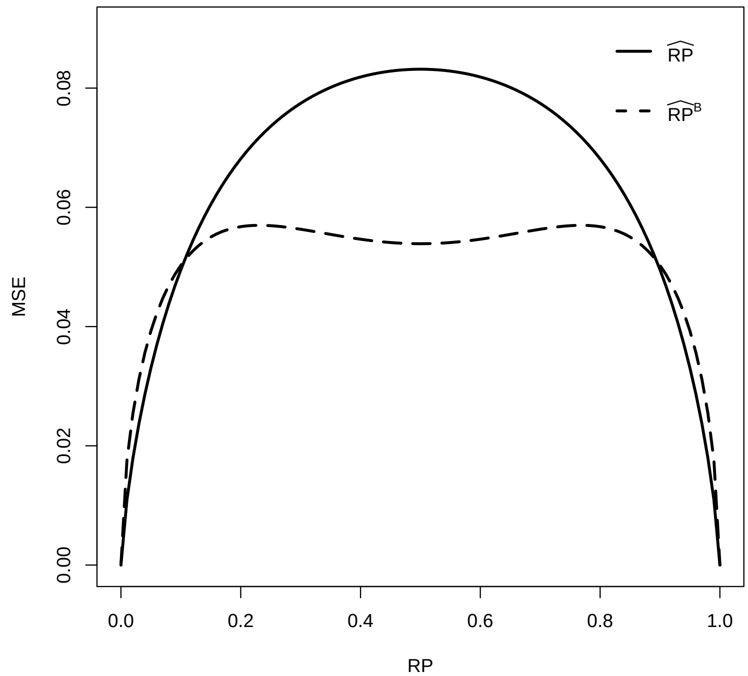
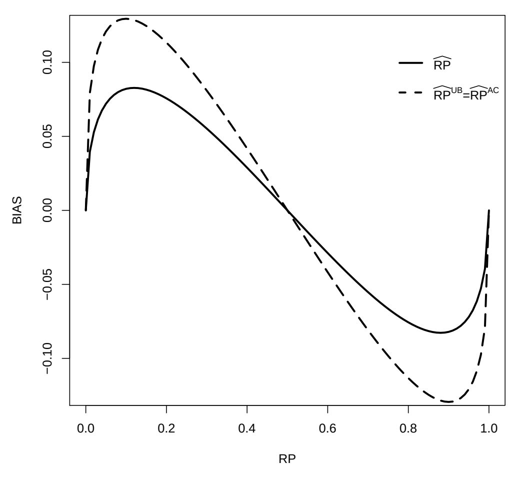
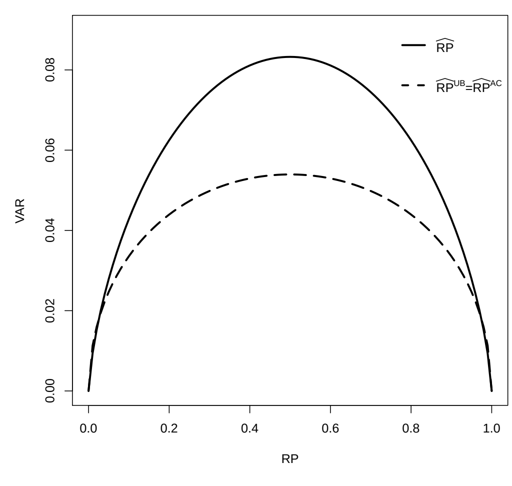


Figure 2: *MSE of $\widehat{RP}^{AC}$ and $\widehat{RP}$ for the Z-test, in function of the RP.*

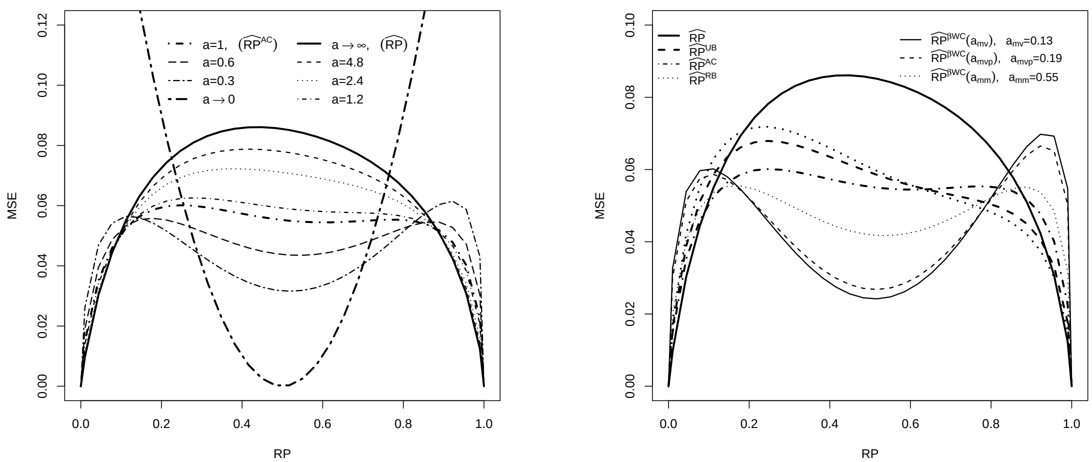


(a) Fig.2a



(b) Fig.2b

Figure 3: *Bias and variance of $\widehat{RP}^{AC}$ and $\widehat{RP}$ for the Z-test, in function of the RP.*



(a) (b)

Figure 4: *MSE of* $\widehat{RP}^{\beta WC}(a)$*, at 8 values of* a*, and of several RP estimators, for the* t*-test with 10 dfs and* $\alpha = 0.05$*, in function of the RP.*

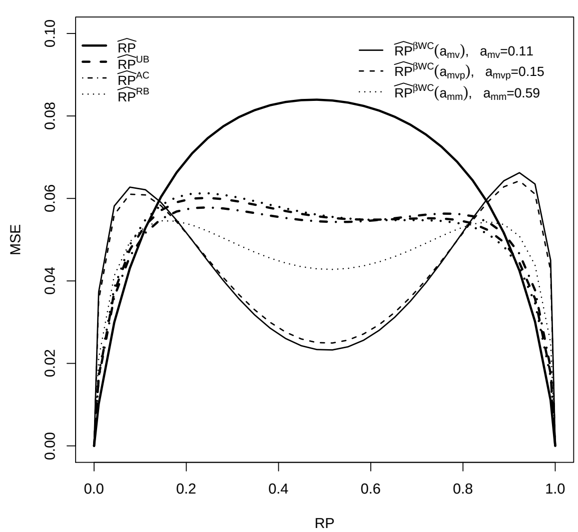


Figure 5: *MSE of several RP estimators for the* t*-test with 30 dfs, in function of the RP.*

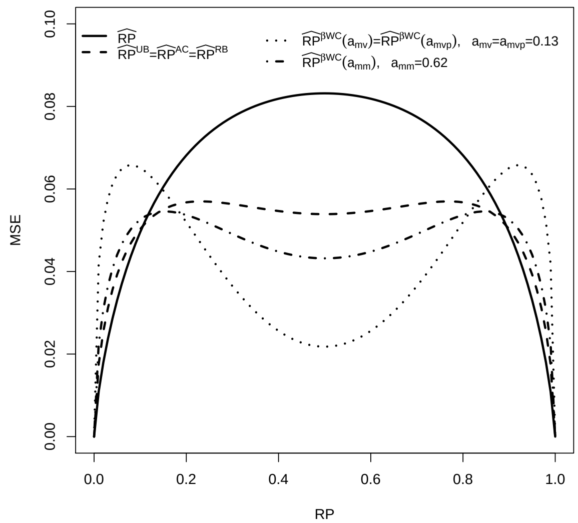


Figure 6: *MSE of several RP estimators for the Z-test, in function of the RP.*

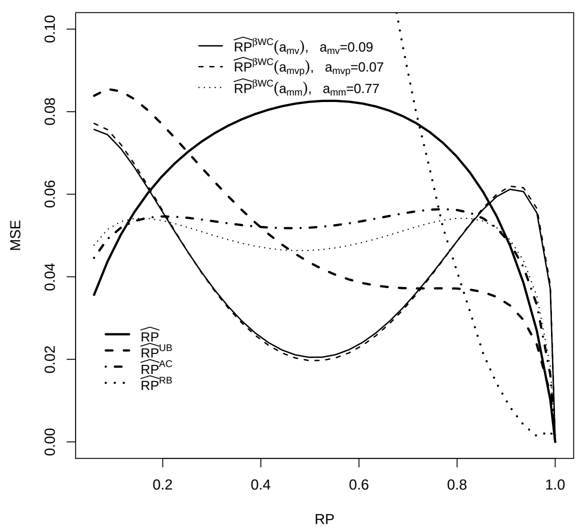


Figure 7: *MSE of several RP estimators for the* χ^2*-test with 30 dfs, in function of the RP* - $\alpha = 0.05$.

						$\widehat{RP}^{\beta WC}(a)$		
η	α	$\widehat{RP}$	$\widehat{RP}^{UB}$	$\widehat{RP}^{AC}$	$\widehat{RP}^{RB}$	a_{mv}	a_{mvp}	a_{mm}
10	0.01	0.00%	18.96%	21.44%	16.98%	33.57%	33.27%	29.04%
	0.05	0.00%	19.01%	21.49%	17.63%	33.52%	33.36%	28.75%
	0.1	0.00%	19.02%	21.52%	17.87%	33.49%	33.35%	28.77%
30	0.01	0.00%	20.58%	21.25%	20.20%	33.45%	33.40%	27.80%
	0.05	0.00%	20.58%	21.26%	20.26%	33.44%	33.39%	27.80%
	0.1	0.00%	20.58%	21.26%	20.29%	33.44%	33.43%	27.46%
50	0.01	0.00%	20.86%	21.21%	20.67%	33.44%	33.43%	27.40%
	0.05	0.00%	20.86%	21.21%	20.69%	33.44%	33.42%	27.40%
	0.1	0.00%	20.86%	21.21%	20.70%	33.44%	33.42%	27.40%
100	0.01	0.00%	21.07%	21.17%	20.98%	33.44%	33.44%	27.01%
	0.05	0.00%	20.86%	21.21%	20.69%	33.44%	33.42%	27.40%
	0.1	0.00%	20.86%	21.21%	20.70%	33.44%	33.42%	27.40%

Table 1: Relative mean gain of RP estimators with respect to $\widehat{RP}$ for the considered values of η and α for the t test.

						$\widehat{RP}^{\beta WC}(a)$		
η	α	$\widehat{RP}$	$\widehat{RP}^{UB}$	$\widehat{RP}^{AC}$	$\widehat{RP}^{RB}$	a_{mv}	a_{mvp}	a_{mm}
10	0.01	0.00%	12.05%	12.72%	10.26%	4.39%	8.34%	13.72%
	0.05	0.00%	11.79%	12.57%	10.66%	4.19%	6.72%	13.44%
	0.1	0.00%	11.66%	12.49%	10.79%	3.68%	6.28%	13.26%
30	0.01	0.00%	12.35%	12.56%	12.06%	4.38%	6.10%	13.76%
	0.05	0.00%	12.31%	12.54%	12.10%	4.22%	5.96%	13.72%
	0.1	0.00%	12.29%	12.53%	12.11%	4.16%	5.05%	13.73%
50	0.01	0.00%	12.45%	12.56%	12.32%	4.62%	5.50%	13.83%
	0.05	0.00%	12.43%	12.55%	12.33%	4.57%	5.44%	13.81%
	0.1	0.00%	12.43%	12.55%	12.33%	4.54%	5.42%	13.80%
100	0.01	0.00%	12.54%	12.57%	12.49%	4.87%	4.87%	13.88%
	0.05	0.00%	12.43%	12.55%	12.33%	4.57%	5.44%	13.81%
	0.1	0.00%	12.43%	12.55%	12.33%	4.54%	5.42%	13.80%

Table 2: Mean relative gain of RP estimators with respect to $\widehat{RP}$ for the considered values of η and α for the t test.

η	α	a_{mv}	a_{mvp}	a_{mm}
10	0.01	0.11	0.21	0.53
	0.05	0.13	0.19	0.55
	0.1	0.13	0.19	0.55
30	0.01	0.11	0.15	0.59
	0.05	0.11	0.15	0.59
	0.1	0.11	0.13	0.61
50	0.01	0.11	0.13	0.61
	0.05	0.11	0.13	0.61
	0.1	0.11	0.13	0.61
100	0.01	0.11	0.11	0.63
	0.05	0.11	0.13	0.61
	0.1	0.11	0.13	0.61

Table 3: Values of parameters a_{mv}, a_{mvp}, and a_{mm} for the t test and for the considered values of α and η.

						$\widehat{RP}^{\beta WC}(a)$		
η	α	$\widehat{RP}$	$\widehat{RP}^{UB}$	$\widehat{RP}^{AC}$	$\widehat{RP}^{RB}$	a_{mv}	a_{mvp}	a_{mm}
4	1,00%	0,00%	19,69%	21,54%	-23,83%	34,01%	33,99%	28,35%
	5,00%	0,00%	18,83%	23,05%	-42,77%	36,13%	36,11%	26,53%
	10,00%	0,00%	18,10%	24,73%	-51,21%	38,53%	38,52%	24,73%
9	1,00%	0,00%	21,25%	21,55%	-164,85%	34,04%	34,02%	28,53%
	5,00%	0,00%	21,63%	23,00%	-191,56%	36,14%	36,12%	26,48%
	10,00%	0,00%	21,52%	24,66%	-192,13%	38,52%	38,50%	24,66%
16	1,00%	0,00%	22,10%	21,53%	-302,77%	34,05%	34,03%	28,51%
	5,00%	0,00%	23,20%	22,96%	-300,29%	36,12%	36,11%	26,61%
	10,00%	0,00%	23,49%	24,62%	-278,45%	38,50%	38,49%	24,62%
30	1,00%	0,00%	22,83%	21,51%	-388,72%	34,05%	34,04%	28,49%
	5,00%	0,00%	24,57%	22,91%	-350,25%	36,10%	36,09%	26,39%
	10,00%	0,00%	25,24%	24,57%	-311,65%	38,49%	38,48%	24,57%

Table 4: Relative mean gain of RP estimators with respect to $\widehat{RP}$ for the considered values of η and α for the χ^2 test.

						$\widehat{RP}^{\beta WC}(a)$		
η	α	$\widehat{RP}$	$\widehat{RP}^{UB}$	$\widehat{RP}^{AC}$	$\widehat{RP}^{RB}$	a_{mv}	a_{mvp}	a_{mm}
4	1,00%	0,00%	8,82%	14,22%	-46,54%	10,13%	9,38%	16,52%
	5,00%	0,00%	11,23%	17,07%	-54,08%	17,11%	16,54%	19,04%
	10,00%	0,00%	13,40%	19,49%	-54,55%	21,85%	21,61%	19,49%
9	1,00%	0,00%	10,95%	14,24%	-225,39%	10,29%	9,55%	16,59%
	5,00%	0,00%	14,07%	17,08%	-214,28%	17,31%	16,75%	19,06%
	10,00%	0,00%	16,60%	19,49%	-195,84%	22,05%	21,57%	19,49%
16	1,00%	0,00%	12,16%	14,25%	-408,81%	10,39%	9,66%	16,61%
	5,00%	0,00%	15,70%	17,07%	-333,37%	17,42%	16,86%	19,15%
	10,00%	0,00%	18,44%	19,47%	-282,98%	22,14%	21,66%	19,47%
30	1,00%	0,00%	13,21%	14,25%	-531,84%	10,49%	9,76%	16,63%
	5,00%	0,00%	17,11%	17,05%	-389,12%	17,49%	16,94%	19,05%
	10,00%	0,00%	20,06%	19,44%	-316,24%	22,19%	21,71%	19,44%

Table 5: Mean relative gain of RP estimators with respect to $\widehat{RP}$ for the considered values of η and α for the χ^2 test.

η	α	a_{mv}	a_{mvp}	a_{mm}
4	0.01	0.1	0.08	0.58
	0.05	0.09	0.07	0.78
	0.1	0.07	0.06	1
9	0.01	0.1	0.08	0.57
	0.05	0.09	0.07	0.78
	0.1	0.07	0.05	0.99
16	0.01	0.1	0.08	0.57
	0.05	0.09	0.07	0.77
	0.1	0.07	0.05	0.99
30	0.01	0.1	0.08	0.57
	0.05	0.09	0.07	0.78
	0.1	0.07	0.05	1

Table 6: Values of parameters a_{mv}, a_{mvp}, and a_{mm} for the χ^2 test and for the considered values of α and η.