Epidemiology Biostatistics and Public Health - 2013, Volume 10, Number 1

BIOSTATISTICS

An Introduction to RP-testing

LUCIO DE CAPITANI[(1)]

Abstract

This paper reviews the concept of Reproducibility Probability and makes a brief introduction to RP-testing. The RP-based version of some common parametric tests is provided. Moreover, particular attention is devoted to the well-known nonparametric Wilcoxon Rank-Sum test. A comparison between the properties of the RP and the p-value is made in order to evaluate the practical utility of these stability indicators. It turns out that the use of the RP to perform statistical tests and to interpret their results, requires more technical analysis, but it provides more interpretable direct information on the stability of the test results.

Keywords: power of a test; p-value; Reproducibility Probability; RP-testing; stability of statistical test outcomes

***DOI:* 10.2427/8756**

1 INTRODUCTION

The reproducibility of a given experiment is a pillar of the Galilean method. Roughly speaking, a scientific theory can be supported by the empirical evidence provided by an experiment only when the latter is reproducible, i.e. if the experiment can be replicated under the same conditions every time it is desired. This requirement is very intuitive: a single, non-reproducible experiment cannot validate a conjecture or a theory. This is due to the fact that all experiments are affected by casualty and it is necessary to be sure that randomness has not played a substantial role in producing the particular observed result. Indeed, if an experiment is reproducible, the impact of randomness vanishes, because previous results can be confirmed or confuted in subsequent identical experiments.

In many situations of practical interest, it is not possible to repeat a particular reproducible experiment many times. For example, a clinical trial performed to assess the benefit of a new drug can be, in theory, replicated, but, is performed only once, or at most, twice since it is very cost-prohibitive both in money and time. In such a situation, a methodology to asses the reproducibility of a single experimental result is essential. A precise methodology can be provided if experiments are evaluated by means of a statistical test. In this case, the reproducibility of the experimental results should be interpreted as the reproducibility of statistical significance, which can be evaluated by computing the Reproducibility Probability (RP) of the test (see [1] and [2]). The RP of a test coincides with the probability of obtaining a rejection of the null hypothesis H_0 (a statistically significant result); its name is due to the fact that it is usually computed when a first experiment produces a significant result, in order to evaluate the probability of obtaining

[(1)] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, via Bicocca degli Arcimboldi n.8, 20126, Milano, Italy. *e-mail*: lucio.decapitani1@unimib.it

statistical significance also in a second, identical, experiment. Naturally, the RP of a test is unknown since, for example, it depends on the true effect of the drug studied in a clinical trial: the greater the effect of the new drug, the greater the probability of obtaining a statistically significant result. Nevertheless, as we will show in detail later, the RP can be estimated using observed data, providing direct information on the stability of the test outcome. Moreover, the RP estimate not only evaluates how much a significant result is reproducible, but it can also play the role of a test statistic. In detail, there exists a general threshold for the statistical significance based on the RP-estimate, that is $1/2$. The decision rule is, therefore: "the null hypothesis is rejected when it is estimated that there is more inclination to reject it than to accept it". This testing technique, hereafter called RP-testing, has been introduced in [8] and it is quite general. The testing rule holds, indeed, for the most commonly used parametric tests, such as the test on one proportion, those on the mean and the variance with normally distributed data, those comparing two proportions, or two means and two variances with normal data. Moreover, it holds, approximately, on some nonparametric tests (one sample Wilcoxon test, Wilcoxon Rank-Sum test, Kendall test). These results make the RP a direct competitor of the p-value. The latter is commonly used in order to perform a test and, at the same time, to evaluate the conflict of the data with the null hypothesis. Moreover, it is commonly thought that a very small p-value corresponds to a highly reproducible significant result but, as clearly explained in [1], the conclusion taken on the base of the p-value is, in general, too optimistic. This is due to the fact that p-value measures how strongly the data contradict the null hypothesis, and not directly the reproducibility of statistical significance.

The aim of this paper is to provide a brief introduction to RP-testing, making also a comparison between the use of RP and p-value. In detail, the paper is organized as follows. In Section 2 we recall some preliminary concepts concerning the general theory of statistical tests. In Section 3 we introduce the concept of Reproducibility Probability and we show how the later can be estimated and used in order to test statistical hypotheses. Section 4 is devoted to the exact RP-testing for some common parametric tests while Section 5 concerns the approximated RP-testing with special emphasis on the Wilconon Rank-Sum test. In Section 6 we compare the RP and the p-value in order to explain why it is advisable to perform and evaluate the stability of a statistical test using the RP-estimates. Finally, Section 7 is devoted to the conclusions.

2 DEFINITION AND PRELIMINARY CONCEPTS

Let X be the random variable (or the random vector) describing a particular feature of a given population and let F denote the probability distribution function of X. A *statistical hypothesis* is an assertion concerning the distribution F of X. In practice, it is usually of interest to discuss two different statistical hypotheses, the *null hypothesis* H_0 and the *alternative hypothesis* H_1. The comparison of these two hypotheses gives rise to a *testing problem* that can be solved by means of a *statistical test*. In detail, a *statistical test* is a rule or procedure, based on a random sample $(X_1, ..., X_n)$ drawn from F, for deciding whether to reject H_0 and, eventually, to accept H_1. In more detail, let $\mathcal{X}$ denote the sample space of observations; that is:

$$\mathcal{X} = \{(x_1, ..., x_n) \in \mathbb{R}^n : (x_1, ..., x_n) \text{ is a possible value of } (X_1, ..., X_n)\} \ .$$

A statistical test is built by defining a subset $\mathcal{C}$ of $\mathcal{X}$, named *critical region*, which leads to the following rule: accept H_0 if, and only if, $(X_1, ..., X_n) \in \mathcal{C}$. Such a decision rule implies two kinds of errors. In detail, the rejection of H_0 when it is true is called a *Type-I error*, and the acceptance of H_0 when it is false is called a *Type-II error*. The definition of the Type-I and Type-II errors provides a useful instrument in order to choose a particular critical region. In detail it is common practice to pre-specify a “desired” level of the Type-I error, usually denoted by α, and then searching an “optimal” critical region among them of level α. A natural optimality criterion is the following: let $\mathcal{C}_\alpha$ and $\mathcal{C}'_\alpha$ be two level-α critical regions and let β and β' be the probability of a Type-II error associated to $\mathcal{C}_\alpha$ and $\mathcal{C}'_\alpha$, respectively. If $\beta > \beta'$ then the critical region $\mathcal{C}'_\alpha$ is better than the critical region $\mathcal{C}_\alpha$. The above criterion leads to the notion of most powerful test or, more generally, to the notion of Uniformly Most Powerful (UMP), Uniformly Most Powerful-Invariant (UMPI), and Uniformly Most Powerful-Unbiased (UMPU) test. A detailed discussion of these topics goes far beyond the scope of the work objectives. We refer the interested reader to [4]. Here, it is worthwhile to note that the above optimality criterion introduces a hierarchy between the compared statistical hypotheses. In particular, the criterion just described puts the null hypothesis on a higher level, since the prescribed level α controls the probability of a false rejection of H_0, while the value of β, even if it is minimized, remains unspecified.

It is common practice to represent a statistical test by means of the so called *critical function*:

$$\Psi_\alpha(X_1, ..., X_n) = \begin{cases} 1 & \text{if} \quad (X_1, ..., X_n) \in \mathcal{C}_\alpha \\ 0 & \quad\quad otherwise \end{cases} \quad . \tag{1}$$

The above representation highlights that a statistical test is a Bernoulli random variable and, then, it underlines the random nature of the statistical test results.

The statistical tests are generally divided into two main groups: *parametric tests* and *non-parametric tests*, where the former are the most widely applied in medical statistics. In the parametric context it is assumed that the random variable X has a parametric distributional model described by the density $f(\cdot;\theta)$. The distributional model is known, but the true value of θ is unknown, and it is of interest to test the following one-sided hypotheses:

$$H_0 : \theta \leq \theta_0 \qquad \text{vs} \qquad H_1 : \theta > \theta_0 \quad . \tag{2}$$

Let $\hat{\theta} = h(X_1, ..., X_n)$ be an estimator of the unknown parameter θ and let K_θ be the distribution function of $\hat{\theta}$:

$$K_\theta(y) = P_\theta(\hat{\theta} < y) \quad .$$

Usually, the estimator $\hat{\theta}$ can be used as test statistic. In detail, let us assume that $\hat{\theta}$ is stochastically increasing in θ and that the *null distribution* K_{θ_0} is known. Moreover, denote by k_q the q-quantile of K_{θ_0}. Under this assumption, it is possible to define the level-α critical region

$$\mathcal{C}_\alpha \equiv \left\{ (x_1, ..., x_n) \in \mathcal{X} : \hat{\theta} > k_{1-\alpha} \right\} \tag{3}$$

which corresponds to the decision rule: “H_0 is rejected if, and only if, the estimate of θ is greater than $k_{1-\alpha}$”. It is easy to verify that the maximum probability of a Type-I error associated to the above critical region is α. Indeed, the probability of Type II error associated to the critical region $\mathcal{C}_\alpha$ depends on θ and it coincides with

$$\beta(\theta) = P_\theta(\hat{\theta} > k_{1-\alpha}) \quad . \tag{4}$$

It is worthwhile to note that, in order to define the critical region $\mathcal{C}_\alpha$, it is only necessary to know the null distribution K_{θ_0}, that is, the knowledge of the non-null distribution is not an essential element in order to define a statistical test. However, if the distribution K_θ is known for all values of θ, it is possible to evaluate, through (4), the magnitude of the probability of the Type-II error and, then, it is possible to evaluate “how good” the test is. In order to evaluate the performance of a statistical test, it is unusual to refer directly to $\beta(\theta)$, and it is more common to use the concept of *power function* of the test:

$$\pi(\theta) = 1 - \beta(\theta) \ .$$

Naturally, the greater the power the better the test. Moreover, it is well-known that, the power function coincides with the expectation of the critical function of the test as a function of θ. In other words, it is the expectation of the test in function of θ:

$$\pi(\theta) = E_\theta[\Psi_\alpha(X_1, ..., X_n)] \ . \tag{5}$$

3 RP ESTIMATION AND TESTING

From (1) and (5) it follows that the statistical test is a Bernoulli random variable with unknown parameter $\pi(\theta)$. This fact highlights that the random nature of the test is completely described by the unknown value of $\pi(\theta)$ and, then, the power is a perfect tool in order to evaluate the variability of the test results. Moreover, the evaluation of the true power of a test is particularly important, since it can be interpreted as RP. Roughly speaking, once a statistical test is computed referring to data from a particular experiment, the true power is the probability of obtaining the same test result in a second, identical experiment. In detail, if we accept H_0 in the first experiment, the probability to accept H_0 even in the second experiment coincides with 1 minus the true power. Otherwise, if we reject H_0 in the first experiment, the probability of a further rejection is the true power itself. The interpretation of the true power in terms of RP is particularly meaningful and clearly shows that the power is not only a technical concept useful in order to define an optimality criterion.

In the following, in order to avoid confusion between the concept of power function and the concept of true power, we will use the terminology used in [1] and we will refer to the true power as the Reproducibility Probability of the test. In detail, let θ^* denote the true value of θ. The Reproducibility Probability of the test is $RP = \pi(\theta^*)$. Naturally, the value RP is unknown since θ^* is unknown too. So, a natural question is: how to estimate it? Since the RP coincides with the parameter of a Bernoulli random variable, it seems natural to estimate it as the proportion of rejections of H_0 in a sequence of repeated tests. However this solution is unfeasible since, in practice, only one test is performed. Another solution (see [2] or [3]) is to start from an estimator of θ and, then, plugging it into the power function. For example, a very natural RP-estimator is

$$\widehat{RP}_\alpha = \pi(\hat{\theta}^\bullet) \ , \tag{6}$$

where $\hat{\theta}^\bullet$ is such that $P\left[\hat{\theta}^\bullet \leq \theta^*\right] = 0.5$. In [3], the RP-estimator (6) is referred to as "50%-lower bound for the RP" since it can be interpreted as the lower bound of a unidirectional confidence interval for the RP at 50% confidence level. However, note that, it is a point estimator since, rougly speaking, it equals the 50%-upper bound. Recently, [3] showed that the reproducibility probability estimators $\widehat{RP}_\alpha$ can also be used for testing hypotheses (2). In particular, under some mild regularity conditions, it is shown that $\widehat{RP}_\alpha > 1/2$ if and only if the null hypothesis is rejected. It then follows that the statistical test described by the critical function

$$\Psi_\alpha(X_1, ..., X_n) = \begin{cases} 1 & \text{if } \widehat{RP}_\alpha > 1/2 \\ 0 & \text{if } \widehat{RP}_\alpha \leq 1/2 \end{cases} \tag{7}$$

is equivalent to the statistical test defined by the critical region (3). In the following we will use the expression "RP-testing" to indicate the usage of the critical function (7) in order to test the hypotheses (2). Moreover, it is well-known that a statistical test can be performed by means of the p-value, denoted by PV, which induces the following critical function

$$\Psi_\alpha(X_1, ..., X_n) = \begin{cases} 1 & \text{if } PV < \alpha \\ 0 & \text{if } PV \geq \alpha \end{cases} \quad . \tag{8}$$

So, under some mild regularity conditions, there are three equivalent ways to perform a parametric statistical test and, in the following, we will discuss the strengths and weaknesses of each one.

For the nonparametric tests, it is harder to demonstrate that it is possible to test one-sided statistical hypotheses through the reproducibility probability estimates. This is due to the fact that it is not possible to define the power function without making some kind of parametric assumptions. A detailed discussion of this topic goes far beyond the scope of the work objectives. However, we will analyze in Section 5.2 the case of the Wilcoxon Rank-Sum test showing that –for this test– it is possible to perform RP-testing, at least asymptotically.

4 EXACT RP-TESTING FOR SOME COMMON PARAMETRIC TESTS

A common problem arising in the analysis of clinical trials is the comparison between two different drugs or treatments. If the effect size of the two treatments is measured by the mean of the variable of interest and it is assumed that the effect of both treatments is normally distributed, then the well-known "Z" and "t" tests for the comparison of two means can be used to detect the most effective treatment. In the following, we briefly recall these two tests with special emphasis on their RP-based version.

4.1 The "Z" test for the comparison of two means when sampling from two normal populations with known variances

Let X and Y be the random variables describing the efficacy of the two treatments. Assume that X and Y are independent normally distributed with (unknown) expected values and variances given by μ_X, μ_Y, σ_X^2, and σ_Y^2, respectively. Moreover, in the following we assume that the variances σ_X^2 and σ_Y^2 are known. To ease the exposition, hereafter we will refer to the treatment described by X as the *placebo*. So, let $X_1, ..., X_m$ be the random variables describing the effect of the placebo on m patients. Similarly, let $Y_1, ..., Y_n$ be the random variables for the n patients under treatment. If μ_X and μ_Y represent, respectively, the effect sizes of the placebo and the treatment, then the statistical hypotheses of interest are

$$H_0: \ \mu_Y \leq \mu_X \qquad \text{vs} \qquad H_1: \ \mu_Y > \mu_X \ . \tag{9}$$

Under the assumptions previously specified, the hypotheses (9) can be verified using the following test

$$\Psi_\alpha(X_1, ..., X_m; Y_1, ...Y_n) = \begin{cases} 1 & \text{if } \ \bar{Y} - \bar{X} > z_{1-\alpha}\sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}} \\ 0 & \text{if } \ \bar{Y} - \bar{X} \leq z_{1-\alpha}\sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}} \end{cases} \tag{10}$$

where α denotes the pre-specified level of the Type-I error probability, $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, and $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution, that is: $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$, where Φ denotes the standard normal distribution function. As it is well-known, the above test can be performed using the p-value which, in this context coincides with

$$PV = 1 - \Phi\left(\frac{\bar{Y}-\bar{X}}{\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}}\right) \tag{11}$$

and the test (10) is equivalent to

$$\Psi_\alpha(X_1,...,X_m;Y_1,...Y_n) = \begin{cases} 1 & \text{if} \quad PV < \alpha \\ 0 & \text{if} \quad PV \geq \alpha \end{cases} \quad . \tag{12}$$

In order to introduce the RP-based version of the test (10) it is necessary to define the power function of the same test. Being $k = \mu_Y - \mu_X$, the power function of the test (10) is

$$\begin{aligned} \pi(k;m,n,\alpha,\sigma_Y^2,\sigma_X^2) &= P_k\left[\bar{Y}-\bar{X} > z_{1-\alpha}\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}\right] \\ &= 1-\Phi\left(z_{1-\alpha} - \frac{k}{\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}}\right) \quad . \end{aligned} \tag{13}$$

In the following we will denote the power function $\pi(k;m,n,\alpha,\sigma_Y^2,\sigma_X^2)$ simply with $\pi(k)$.

Now, it is worthwhile to observe that the estimator $\hat{k} = (\bar{Y} - \bar{X})$ is unbiased for $k = \mu_Y - \mu_X$ and it is symmetrically distributed (in detail, it is normally distributed). Then $\hat{k}^\bullet \equiv \hat{k}$ and, following [3], the RP estimator

$$\widehat{RP}_\alpha = 1-\Phi\left(z_{1-\alpha} - \frac{\hat{k}}{\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}}\right) \quad . \tag{14}$$

can be used to define the following test

$$\Psi_\alpha(X_1,...,X_m;Y_1,...Y_n) = \begin{cases} 1 & \text{if} \quad \widehat{RP}_\alpha > 1/2 \\ 0 & \text{if} \quad \widehat{RP}_\alpha \leq 1/2 \end{cases} \tag{15}$$

which is equivalent to test (10) and (12). To clearly see that the test (15) is equivalent to test (10) observe that, as shown in Figure 1, the power function $\pi(k)$ is strictly increasing in k. Moreover, when k is equal to the critical value $z_{1-\alpha}\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}$, $\pi(k)$ equals 1/2:

$$\pi\left(z_{1-\alpha}\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}\right) = 1/2 \quad .$$

Therefore, $\widehat{RP}_\alpha > 1/2$ if, and only if, $\bar{Y}-\bar{X} > z_{1-\alpha}\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}$.

It is worthwhile to note that between the p-value and the RP of the test (10) there exists a 1 to 1 correspondence. In detail, from definition (11) it follows that:

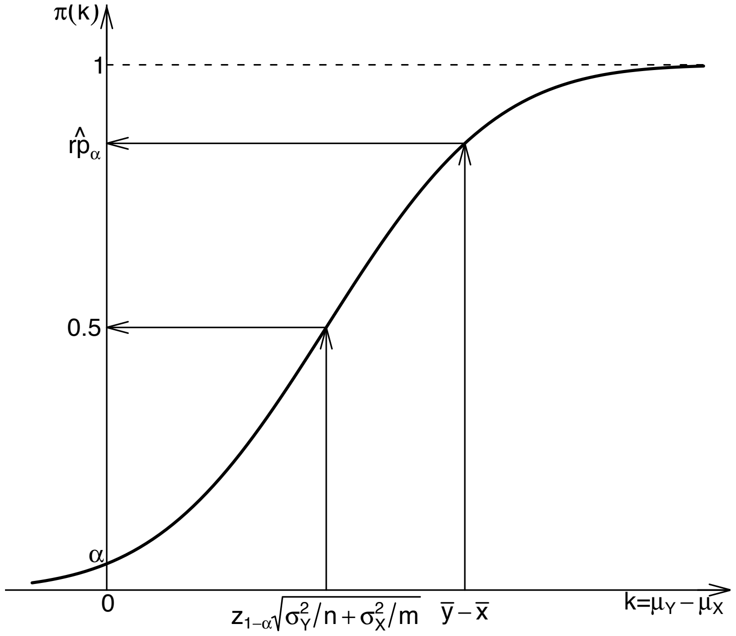
$$\frac{\hat{k}}{\sqrt{\frac{\sigma_Y^2}{n}+\frac{\sigma_X^2}{m}}} = \Phi^{-1}(1-PV) \quad . \tag{16}$$

Consequently:

$$\widehat{RP}_\alpha = 1-\Phi\left[z_{1-\alpha} - \Phi^{-1}(1-PV)\right] \quad . \tag{17}$$

In Figure 2 the plot of $\widehat{RP}_\alpha$ as a function of p is shown. Moreover, in Figure 3 the difference between the p-value and $\widehat{RP}_\alpha$ is graphically displayed.

Figure 1. Power function of the test (10) when $m = n = 10$, $\sigma_X^2 = \sigma_Y^2 = 1$ and $\alpha = 0.05$. The plot emphasizes that the test (15) is equivalent to the test (10) since it clearly shows that $\widehat{RP}_\alpha > 1/2$ if and only if $\bar{Y} - \bar{X} > z_{1-\alpha}\sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}}$. Moreover, the plot illustrates how the estimator $\widehat{RP}_\alpha$ is defined by the simple plug-in method.



4.2 The "t" test for the comparison of two means when sampling from two normal populations with common unknown variance

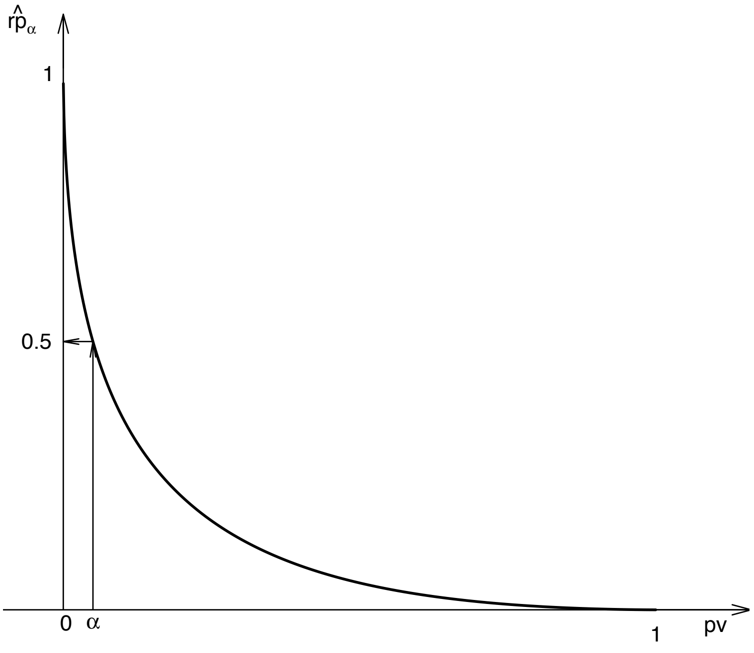
Let X and Y be two independent and normally distributed random variables with (unknown) expected value and (unknown) common variance given by μ_X, μ_Y, σ^2, respectively. In order to test hypotheses (9), it is possible to use the well-known "t" test:

$$\Psi_\alpha(X_1,...,X_m;Y_1,...,Y_n) = \begin{cases} 1 & \text{if} \quad \bar{Y} - \bar{X} > t_{1-\alpha}(m+n-2)\sqrt{S_p^2\left(\frac{m+n}{mn}\right)} \\ 0 & \text{if} \quad \bar{Y} - \bar{X} \leq t_{1-\alpha}(m+n-2)\sqrt{S_p^2\left(\frac{m+n}{mn}\right)} \end{cases} \tag{18}$$

where $t_{1-\alpha}(\nu)$ denotes the $(1-\alpha)$-quantile of the Student's t distribution with ν degrees of freedom and S_p^2 denotes the pooled estimator of the common variance:

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{m+n-2} \quad .$$

Figure 2. Behavior of the RP estimator (14) as a function of the p-value.



In this context the p-value coincides with

$$PV = 1 - \mathcal{T}_{(m+n-2)}\left(\frac{\bar{Y}-\bar{X}}{\sqrt{S_p^2\left(\frac{m+n}{mn}\right)}}\right) \tag{19}$$

where $\mathcal{T}_\nu$ denotes the Student's t distribution function with ν degrees of freedom. Naturally, the p-value defined above, can be used to define a test equivalent to (18) using the well-known decision rule defined in (8). In order to define the power function of the test (18), we recall that

$$\frac{\bar{Y}-\bar{X}}{\sqrt{S_p^2\left(\frac{m+n}{mn}\right)}} \tag{20}$$

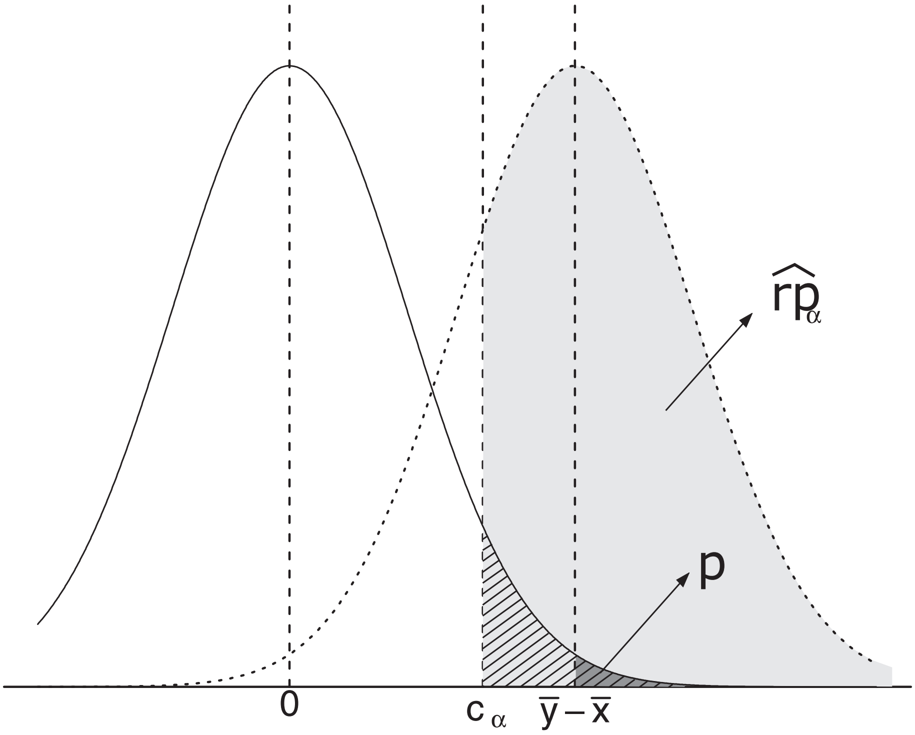
follows the non-central t distribution with $(m+n-2)$ degrees of freedom and non-centrality parameter given by

$$\left(\frac{mn}{m+n}\right)^{1/2} k \qquad \text{with} \qquad k = \frac{\mu_Y - \mu_X}{\sigma} \quad .$$

Let $\mathcal{T}_\nu^\delta$ denote the non-central t distribution with ν degrees of freedom and non-centrality parameter δ. Then, the power function is given by:

$$\pi(k) = 1 - \mathcal{T}_{m+n-2}^{\left(\frac{mn}{m+n}\right)^{1/2}k}(t_{1-\alpha}(m+n-2)) \tag{21}$$

Figure 3. Graphical representation of the pv value and $\widehat{rp}_\alpha$. The solid curve represents the null distribution while the dotted curve coincides with the density of $\bar{Y}-\bar{X}$ when $\mu_Y-\mu_X=\bar{y}-\bar{x}$. The estimates of the RP is the light gray area while the p-value coincides with the dark gray one. In the graph, c_α denotes the critical value and the dashed area equals α.



The non-central t distribution is not symmetric and, consequently, the RP-estimator $\widehat{RP}_\alpha$ is not defined by the simple plug-in method as in the previous section. In detail, to calculate the estimator $\widehat{RP}_\alpha$ it is necessary to obtain the point estimator $\hat{k}^\bullet$ which is implicitly defined as the solution of the equation

$$\hat{k}^\bullet\ :\ t_{1/2}^{\left(\frac{mn}{m+n}\right)^{1/2}\hat{k}^\bullet}(m+n-2)=\left(\frac{mn}{m+n}\right)^{1/2}\hat{k}$$

where $t_q^\delta(\nu)$ denotes the q-quantile of the non-central t distribution with ν degrees of freedom and non-centrality parameter δ. The above equation can not be solved analytically but it is easily solved numerically. Again, starting from the estimator $\widehat{RP}_\alpha=\pi(\hat{k}^\bullet)$ it is possible to define the test (7) which, following the results in [3], turns out to be equivalent to test (18).

5 SOME EXAMPLE OF APPROXIMATED RP-TESTING

In several applications, the assumption of normality of the parent distributions would be inappropriate. Furthermore, there are situations of practical relevance where it is not possible to advance any assumption on the distributional model for X and Y. In these cases, the problem of comparing the effects of the treatments described by X and Y, can be solved using, among others, the well-known asymptotic "Z" test for the comparison of two means, or the widely applied nonparametric Wilcoxon Rank-Sum (WRS) test. In the following two subsections, we will show how to define the RP-based version of these two tests. It is worthwhile to note that, in these two cases, the RP-based test is not exact (i.e. its level is only approximately α). In detail, it is possible to replicate the asymptotic "Z" test for the comparison of two means with an RP-based test but the latter is not exact since also the first is approximate. A different scenario arises when analyzing the WRS test. In fact, in this case the exact WRS test is well defined since the null distribution of the WRS statistic is well-known. On the contrary, the exact distribution of the WRS statistic can not be, in general, derived under the alternative hypothesis and, consequently, the exact power of the test can not be defined. As a consequence, it is possible to perform the exact WRS using the critical value or the p-value, but no exact RP-based test can be defined. However, as outlined in section 4.2, the WRS test can be approximately replicated using an RP-estimator. Moreover, in Section 4.2 we will show that the asymptotic RP-based test approximates the exact WRS test very well.

5.1 The asymptotic "Z" test for the comparison of two means when sampling from two arbitrary distributions with unknown, finite, variances

If X and Y can not be retained Gaussian and it is not realistic to assume that $\sigma_X^2 = \sigma_Y^2$, the testing problem (9) is usually solved assuming that σ_X^2 and σ_Y^2 are finite and, then, using the following result:

$$\frac{(\bar{Y}-\bar{X})-(\mu_Y-\mu_X)}{\sqrt{\frac{S_Y^2}{n}+\frac{S_X^2}{m}}} \overset{a}{\sim} \mathcal{N}(0,1) \quad . \tag{22}$$

In the previous expression $\overset{a}{\sim}$ stands for "asymptotically distributed". Then, if n and m are large enough to justify the application of the asymptotic approximation (22), the following asymptotic test can be used:

$$\Psi_\alpha(X_1,...,X_m;Y_1,...Y_n) = \begin{cases} 1 & \text{if} \quad \bar{Y}-\bar{X} > z_{1-\alpha}\sqrt{\frac{S_Y^2}{n}+\frac{S_X^2}{m}} \\ 0 & \text{if} \quad \bar{Y}-\bar{X} \leq z_{1-\alpha}\sqrt{\frac{S_Y^2}{n}+\frac{S_X^2}{m}} \end{cases} \quad . \tag{23}$$

Naturally, the actual level α^* of the above test is different from the nominal level α. However, the difference between α and α^* decreases as n and m increases and it becomes negligible if n and m are sufficiently large (it is common practice to retain that the test (23) can be applied when $n \geq 50$ and $m \geq 50$).

In analogy with Section 4.1, it is possible to define the RP-estimator

$$\widehat{RP}_\alpha = 1 - \Phi\left(z_{1-\alpha} - \frac{\bar{Y}-\bar{X}}{\sqrt{\frac{S_Y^2}{n}+\frac{S_X^2}{m}}}\right) \tag{24}$$

which defines, through (15), a test equivalent to (23).

5.2 Approximated RP-testing for the Wilcoxon Rank-Sum test

The WRS test is a very widely used nonparametric test for comparing the distributions F and G of the continuous random variables X and Y, respectively. The testing problem solved by the WRS test is often presented as follows (see, e.g. [4]):

$$H_0 : Y \overset{d}{=} X \qquad \text{vs} \qquad H_1 : Y >_{st} X \quad , \tag{25}$$

where "$\overset{d}{=}$" means "equality in distribution", and "$>_{st}$" stands for "stochastically strictly larger in the sense of the usual stochastic ordering". Starting from the independent random samples $(X_1, ..., X_m)$ and $(Y_1, ..., Y_n)$, the test statistic is:

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij} \qquad \text{where} \qquad I_{ij} = \begin{cases} 1 & \text{if } X_i < Y_j \\ 0 & \text{otherwise} \end{cases} .$$

Note that W is strictly related to the following estimator of $p_1 = P_{F,G}(X < Y)$:

$$\hat{p}_1 = \frac{W}{mn} .$$

Moreover, it is worth noting that this is not a test on the difference between the medians of two populations, although sometimes it is referred to in this way.

Let α be the Type-I error probability and let $w_{1-\alpha}$ denote the $(1-\alpha)$-quantile of the exact null distribution of W, which can be exactly calculated. Then, the WRS test is:

$$\Psi_\alpha(X_1, ..., X_m; Y_1, ..., Y_n) = \begin{cases} 1 & \text{if} \quad W > w_{1-\alpha} \\ 0 & \text{if} \quad W \leq w_{1-\alpha} \end{cases} . \tag{26}$$

The power of the test is, hence:

$$\pi(F, G; m, n, \alpha) = P_{F,G}(W > w_{1-\alpha}) .$$

We recently showed that (see [5]), under certain quite general conditions on the distributions F and G and under the knowledge of F, there exists an RP-estimator assuring that the RP-based test is equivalent to the classical WRS test (26). Since this result is merely theoretical (because in practice F is unknown) applied RP-testing needs an approximation of the power function to derive, through the plug-in principle, the related RP-estimator.

An approximation for the power is derived from the asymptotic normality of W. In detail, being $p_2 = P(X < Y \wedge X < Y')$ (Y' and Y i.i.d.), and $p_3 = P(X < Y \wedge X' < Y)$ (X' and X i.i.d.), in [4] it is shown that, when n and m diverge together, we have:

$$\frac{W - mnp_1}{\sqrt{V(p_1, p_2, p_3)}} \overset{a}{\sim} \mathcal{N}(0, 1) \tag{27}$$

where $V(p_1, p_2, p_3) = mnp_1(1-p_1) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2)$ is the variance of W. The limit distribution (27) can be used to define several estimators of $\pi(F, G; m, n, \alpha)$, that, in practice, represent RP-estimators that might be adopted to perform RP-testing. Some of these estimators have been recently presented in [6] and they have been applied to RP-testing, showing very good performances. In more detail, directly from (27) it turns out that

$$\pi(F, G; m, n, \alpha) \approx \Phi\left[\frac{mn\left(p_1 - \frac{1}{2}\right) - z_{1-\alpha}\sqrt{\frac{mn(N+1)}{12}}}{\sqrt{V(p_1, p_2, p_3)}}\right] . \tag{28}$$

From expression (28) the following RP-estimator can be introduced:

$$\widehat{RP}_\alpha = \Phi\left[\frac{mn\left(\hat{p}_1 - \frac{1}{2}\right) - z_{1-\alpha}\sqrt{\frac{mn(N+1)}{12}}}{\sqrt{V(\hat{p}_1, \hat{p}_2, \hat{p}_3)}}\right] \tag{29}$$

where

$$\hat{p}_2 = \frac{1}{n^2 m} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{n} I_{ij} I_{ik} \,, \qquad \hat{p}_3 = \frac{1}{nm^2} \sum_{i=1}^{m} \sum_{k=1}^{m} \sum_{j=1}^{n} I_{ij} I_{kj} \;\; .$$

Several RP-estimators can be obtained from particular simplifications of (28). The latter lead to power approximations that depend only on p_1. The most known simplification is due to [7], where it is assumed that the difference between F and G is quite small, so that the variance of W can be approximated by its value under H_0, i.e. $V(p_1, p_2, p_3) \approx mn(m+n+1)/12$. Being $N = m + n$, the resulting power approximation is:

$$\pi_{\mathbf{n},\alpha}(F,G) \approx \Phi \left[\sqrt{\frac{12mn}{N+1}} \left(p_1 - \frac{1}{2} \right) - z_{1-\alpha} \right] \tag{30}$$

and the related RP-estimator becomes:

$$\widehat{RP}_\alpha = \Phi \left[\sqrt{\frac{12mn}{N+1}} \left(\hat{p}_1 - \frac{1}{2} \right) - z_{1-\alpha} \right] \;\; . \tag{31}$$

Another RP-estimator performing well is derived from a result provided by [8], who obtained bounds for the variance of W in function of p_1. In particular, they showed that $V(p_1, p_2, p_3) \leq V_U(p_1)$, where

$$V_U(p_1) = \begin{cases} U(p_1, m, n) & if \;\; \frac{1}{2} \leq p_1 \leq 1 \\ U(1-p_1, n, m) & if \;\; 0 \leq p_1 < \frac{1}{2} \end{cases}$$

and

$$U(p, m, n) = mn \left[v \left(\frac{k}{3} - (1-p)^2 \right) + u \left(-\frac{2k}{3} + 1 - p^2 \right) + \frac{k}{3} - p(1-p) \right]$$

with $u = \min(m, n)$, $v = max(m, n)$ and $k = 1 - (2p-1)^3/2$. The related estimator is, hence:

$$\widehat{RP}_\alpha = \Phi \left[\frac{mn\left(\hat{p}_1 - \frac{1}{2}\right) - z_{1-\alpha} \sqrt{\frac{mn(N+1)}{12}}}{\sqrt{V_U\left(\hat{p}_1\right)}} \right] \;\; . \tag{32}$$

Other RP-estimators can be found in [6]. When these RP-estimators are adopted for RP-testing, the resulting RP-based test (15) does not exactly correspond to the classical, exact, WRS test. In order to evaluate the discrepancy between these two tests, it is useful to introduce the contingency table described in Table 1, which represents the joint distribution of the WRS test $\Psi_\alpha(W)$ and its RP-based version $\Psi_\alpha\left(\widehat{RP}_\alpha\right)$.

The outcomes $(1, 0)$ and $(0, 1)$ represent the possible disagreement between the tests, and have probability ϵ_1 and ϵ_2, respectively. Note that, under the alternative hypothesis, the powers of the classical WRS test and of the RP-based test are $E\left[\Psi_\alpha(W)\right] = \pi$ and $E[\Psi_\alpha(\widehat{RP}_\alpha)] = \pi'$, respectively. On the contrary, under the null hypothesis, π and π' coincides with the actual level of $\Psi_\alpha(W)$ and $\Psi_\alpha\left(\widehat{RP}_\alpha\right)$. In [6] a wide simulation study is performed to evaluate the performances of several RP-estimators obtaining that the RP-estimator that performs better, both in terms of disagreement and Mean Square Error (MSE), is (32). In order to give an idea of the possible disagreement between the WRS test and the RP-based test build starting from estimator (32), let us consider the following scenario:

Table 1. Joint distribution of the WRS test $\Psi_\alpha(W)$ and its RP-based version $\Psi_\alpha\left(\widehat{RP}_\alpha\right)$

	$\Psi_\alpha\left(\widehat{RP}_\alpha\right)$		
$\Psi_\alpha(W)$	0	1	
0	$1-\pi-\epsilon_1$	ϵ_1	$1-\pi$
1	ϵ_2	$\pi-\epsilon_2$	π
	$1-\pi'$	π'	

Table 2. Joint distribution of $\Psi_\alpha(W)$ and $\Psi_\alpha\left(\widehat{RP}_\alpha\right)$ when $k=1$. In this scenario the percentage disagreement is 0.046%

	$\Psi_\alpha\left(\widehat{RP}_\alpha\right)$		
$\Psi_\alpha(W)$	0	1	
0	0.95050	0.00046	0.95096
1	0.00000	0.04904	0.04904
	0.95050	0.04950	1.00000

- X follows the exponential distribution with parameter $\theta=1$ while Y follows the exponential distribution with parameter $\theta=k\geq 1$. In this case we have that $Y\overset{d}{=}kX$.
- we choose the following 3 different values for k: $k=1$, which correspond to the null hypothesis of equality in distribution; $k=1.45$; $k=1.9$;
- $m=n=60$;
- $\alpha=0.05$

In a Monte-Carlo study with 100.000 replications we obtain the contingency tables reported in Table 2-4 which describe the simulated joint distribution of $\Psi_\alpha\left(\widehat{RP}_\alpha\right)$ and $\Psi_\alpha(W)$ for the 3 different values of k.

As it can be observed from the above tables, the probability of disagreement is low enough to be considered negligible. A very similar result can be found considering different scenarios as demonstrated by the simulations performed in [6], where the overall percentage disagreement turns out to be 0.15%.

Table 3. Joint distribution of $\Psi_\alpha(W)$ and $\Psi_\alpha\left(\widehat{RP}_\alpha\right)$ when $k=1.45$. In this scenario the percentage disagreement is 0.207%

	$\Psi_\alpha\left(\widehat{RP}_\alpha\right)$		
$\Psi_\alpha(W)$	0	1	
0	0.46395	0.00207	0.46602
1	0.00000	0.53398	0.53398
	0.46395	0.53605	1.00000

Table 4. Joint distribution of $\Psi_\alpha(W)$ and $\Psi_\alpha\left(\widehat{RP}_\alpha\right)$ when $k = 1.90$. In this scenario the percentage disagreement is 0.085%

	$\Psi_\alpha\left(\widehat{RP}_\alpha\right)$		
$\Psi_\alpha(W)$	0	1	
0	0.08958	0.00085	0.09043
1	0.00000	0.90957	0.90957
	0.08958	0.91042	1.00000

6 DISCUSSION: PROS AND COS OF P-VALUE AND RP

As mentioned in Section 1 and Section 2, a statistical test is a Bernoulli random variable with parameter equals to the RP. This observation is relevant for two reasons:

1. it underlines that the results obtained from a statistical test are random;
2. it explains that the probabilistic features of a statistical test are governed entirely by the RP, which is the only parameter of the Bernoulli random variable. For example, the expected value of the test coincides with the RP while its variance is given by $RP(1-RP)$.

From observation 1. it turns out that, when evaluating the results of an experiment by means of a statistical test, it would be desirable to join the test result with a stability indicator which reflects the variability of the test itself. This is the reason why tests are usually performed using the p-value instead of the test statistic. In fact, the p-value measures the evidence against or in favor to H_0 and is, then, useful to understand how stable the test result is. Moreover, from observation 2., it is clear that the RP is the natural tool for evaluating the stability of the test results, since it is the only parameter of the test. Furthermore, as shown in the previous sections, the RP can be also used to perform the test as well as the p-value.

If possible, we advise using the RP in place of the p-value both to perform the statistical test and to interpret its results. To thoroughly motivate our point of view, in the following we give a point by point description of the pros and cons of the p-value and the RP.

Pros of the p-value

- The test based on it is equivalent to the original, both in the parametric and in the non-parametric contexts.
- It is easy to compute, since its calculation refers to the null distribution of the test statistic, which is known.

Cons of the p-value

- It is hard to interpret, and this leads to many misinterpretations; mainly, it is often confused with the Type-I error probability or with an estimate of the latter.
- It reports only indirectly the stability of the statistical test outcome; moreover, it can lead to overly optimistic interpretations of the stability (see Goodman,1992). For example, from the relation (17) it turns out that for the "Z" test described in Section 4.1 a p-value of 3% towards an α of 5% corresponds to an estimate of the RP of, just, 59%.

Pros of the RP

- It directly reports the stability of the test outcome (i.e. the estimate of the reproducibility probability).
- It is easy to interpret.
- The adoption of the RP-estimate implies that the Type-I error probability α should be set before analyzing data, discouraging ex-post adjustments of α.
- The RP-estimate allows to discriminate significant results more than the p-value as clearly shown in Figure 2.
- In addition to pointwise estimate of the RP, confidence interval for the RP can be computed. Then, the adoption of the RP perspective brings naturally the possibility of evaluating if only plausible values of the RP are statistically significant, this is a sort of confidence interval on the statistically significant results.

Cos of the RP

- The test based on the RP estimate is, in the nonparametric context, just an approximated version of the original one. Nevertheless, the results obtained in the context of the WRS test show that, for the latter test, the approximation is very good.
- In order to obtain the RP estimate, in the parametric framework it is necessary to refer to the estimated distribution of the test statistic under the alternative hypothesis (e.g. a non-central Student's t) and to its inverse, making the RP-computation more complicated than that of the p-value.
- In order to obtain the RP estimate in the nonparametric framework it is necessary to resort to asymptotic approximation of the distribution of the test statistic under the alternative hypothesis, or, perhaps more simply, to resort to computationally intensive methods, such as the Monte Carlo method (see [6], for details).

The above arguments emphasize that the adoption of the RP to perform and evaluate the stability of a statistical test, requires, undoubtedly, more technical analysis with respect to that required for the usage of the p-value, especially in the nonparametric context. However, it is clear that the RP is much better interpretable than the p-value and it gives to the researcher a direct and excellent instrument to evaluate the stability of the test results. In our opinion, this fully justifies the technical effort required for handling the RP.

7 CONCLUSIONS

The estimation of the reproducibility probability of a test is a key concept in the theory of statistical testing, since it is a useful instrument to perform the test itself and, moreover, to interpret its results. The RP is very interpretable and it sheds light on the stability of the test directly, in contrast to the p-value, which assesses the stability only indirectly and is so difficult to interpret. Moreover, the RP defines a user-friendly decision rule such as the p-value. The use of the RP requires more technicalities with respect to the use of the p-value but the information provided by the first indicator is quite better than those given by the latter. In the parametric context, the RP-based version of almost all the commonly used tests can be obtained with little effort. The non-parametric context requires more attention since, in general, it is not possible to perform RP-testing exactly but only asymptotically. However, the results obtained for the WRS test, are encouraging since simulations show that the disagreement between the exact WRS test and the (asymptotic) RP-based test is negligible even for small sample sizes.

ACKNOWLEDGMENTS: Lucio De Capitani has done this work during a research fellowship co-funded by "Dote Ricercatori", FSE-Lombardia. Partial support was also provided by the Italian Ministry of Research (MIUR) (protocol 2007 AYHZWC, Statistical Methods for Learning in Clinical Research). Thanks are due to Daniele De Martini, whose comments and suggestions helped improve the manuscript.

References

[1] Goodman SN. A comment on replication, p-values and evidence. Statistics in Medicine 1992; 11: 875-879

[2] Shao J, Chow SC. Reproducibility Probability in Clinical Trials. Statistics in Medicine 2002; 21: 1727-1742

[3] De Martini D. Reproducibility Probability Estimation for Testing Statistical Hypotheses. Statistics and Probability Letters 2008; 78: 1056-1061.

[4] Lehmann EL, Romano JP. Testing Statistical Hypotheses. New York: Springer, 2005

[5] 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: 937-946

[6] De Capitani L, De Martini D. Reproducibility Probability Estimation and Testing for the Wilcoxon rank-sum test. Quaderni del dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali dell'Univertistà degli studi di Milano-Bicocca 2010; n. 191

[7] Noether GE. Sample size determination for some common non-parametric tests. Journal of the American Statistical Association 1987; 82: 645-647

[8] Birnbaum ZW, Klose OM. Bounds for the variance of the Mann-Whitney statistic. Annals of Mathematical Statistics 1957; 28, 933-945