

The Emotionality of the p-Value, the Rationality of the Post-Hoc Power: How to Interpret the Results of Statistical Tests

*L'Emotività del Valore-p, la Razionalità della Potenza Post-Hoc:
Come Interpretare i Risultati dei Test Statistici*

Daniele De Martini

Dipartimento SEMEQ, Via Perrone 18, 28100 – Novara, demartin@eco.unipmn.it

Riassunto: Recentemente diversi Editor di riviste di scienze applicate hanno iniziato a chiedere, come completamento dei risultati statistici forniti dai test d'ipotesi, la cosiddetta potenza post-hoc. Ciò denota l'aver intuito l'importanza di questo indicatore, che tuttavia non ha ancora avuto adeguata razionalizzazione. Questo lavoro si propone di fornire un'adeguata spiegazione razionale dell'importanza della potenza post-hoc come indicatore della “stabilità” di una decisione statistica, anche in relazione all'altro tradizionale indicatore: il p-value.

Keywords: Stability of Statistical Decisions, Reproducibility, Consciousness, Power Estimators, Plug-In Estimators, Consistency

1. Discussion

1.1 Statistical tests are a random decision procedure. In this study we will refer to them as one-sample tests without loss of generality.

1.2 In the spirit of Lehmann (1997) let us consider the i.i.d. r.v. $X_1, \dots, X_n$, where X_i has distribution F and values in E . F belongs to the space $P = P_0 \cup P_1$, where $P_0 \cap P_1 = \emptyset$. Let us also consider the hypotheses $H_0: F \in P_0$, and $H_1: F \in P_1$. We define the Acceptance region A_n and the Rejection region R_n , where $E^{\otimes n} = A_n \cup R_n$ and $A_n \cap R_n = \emptyset$, such that, when the Type I error $\alpha \in (0, 1)$ is specified, we obtain $\Pr((X_1, \dots, X_n) \in R_n / F \in P_0) \leq \alpha$.

1.3 The statistical test Φ_α takes values in $\{0, 1\}$, standing for Acceptance/Rejection of H_0 , and it results:

$$\Phi_\alpha(X_1, \dots, X_n) = \begin{cases} 0 & \Leftrightarrow (X_1, \dots, X_n) \in A_n \\ 1 & \Leftrightarrow (X_1, \dots, X_n) \in R_n \end{cases}$$

1.4 α is the parameter of the test, that is, a prefixed and low probability threshold (usually 1% or 5%). This defines, through R_n , the set of observations such that, if one considers H_0 to be true, are of very low total probability (maximum probability of α). Hence, if one of these observations would be realized, we are led to reject the hypotheses on which the calculations of the probability are based, i.e. H_0 .

1.5 Now let us define R_n^F as the boundary of R_n . We observe that, in varying α , the rejection regions are nested, and they shrink when α decreases.

1.6 After having observed the data $(x_1, \dots, x_n)$, it is usual to compute the p-value of the test, i.e. the type I error $a_n(ob)$ that would have to be defined to have a boundary result, that is $(x_1, \dots, x_n) \in R_n^F$.

1.7 Consequently, we have:

$$\Phi_\alpha(X_1, \dots, X_n) = \begin{cases} 0 \Leftrightarrow a_n > \alpha \\ 1 \Leftrightarrow a_n \leq \alpha \end{cases}$$

which is considered as an alternative definition of the test based on the random variable a_n . Thus, $a_n(ob)$ is useful for calculating the success of the test.

1.8 Furthermore, the p-value is an indicator of the “stability” of the decision: $\Phi_\alpha(x_1, \dots, x_n)$. In fact, where $\Phi_\alpha=1$, the lower the value $a_n(ob)$, the more the observations are nested in a rejection region whose boundary is far away from R_n^F . That is, the further the result is from the threshold of probability α (in which jumping over would change the decision), the more stable the random decision. On the other hand, in the case of $\Phi_\alpha=0$, the higher the result $a_n(ob)$, the more stable the decision.

1.9 The use of the p-value is absolutely diffuse and generalized, but it is not logical in this: *in the moment that a decision is taken it is not very useful to calculate how far one is from making the opposite decision*. The behavior is in this case one of *emotionality*.

1.10 On the other hand, $\Phi_\alpha(X_1, \dots, X_n)$ is a Bernoulli random variable with values in $\{0, 1\}$, whose stability is that of being seen in terms of “non-variability”; the variability to read itself in terms of mathematical variance, that for a Bernoullian variable is equal to $\pi_n(1-\pi_n)$, where $\pi_n = E[\Phi_\alpha(X_1, \dots, X_n)]$, that is, the power of the test.

1.11 Therefore, from the parabolic form of $\pi_n(1-\pi_n)$, the closer π_n is to 0 or 1, the more stable the test. However, π_n is unknown, but it can be estimated from an appropriate estimator p_n . Thus, the result of the test in terms of stability is to be interpreted as follows: *the closer $p_n(ob)$ is to 0 or 1, the less the estimated variance is, then the test (i.e. the decision outcome) is estimated to be more stable*.

1.12 $p_n(x_1, \dots, x_n) = p_n(ob)$ is thus to be considered as a *rational* indicator of the stability of the statistical decision. In the literature, this indicator is called *Post-Hoc Power* (PHP).

1.13 From the probabilistic point of view, a_n under H_0 has uniform distribution on $(0, 1)$, while under H_1 it has an asymmetric distribution and tends to 0 in probability as n tends to ∞ . In regards to π_n , there are various appropriate estimators; in the parametric field they are usually based on the estimate of the non centrality parameter (see Rukhin, 1993) to be used as the argument of the power function (Cohen 1988) for obtaining the power estimate; in the nonparametric sphere, they can be used asymptotic approximations of the power functions, or else the estimator based on the plug-in (Collings and Hamilton, 1998), which is simple and generalizable. Indeed, let F_n be the empirical df provided by $X_1, \dots, X_n$, and let $X^*_1, \dots, X^*_n$ be a sample from F_n ; then $p_n = E[\Phi_\alpha(X^*_1, \dots, X^*_n) | F_n]$ is the plug-in estimator, which results in being a V-statistic (De Martini, 2004). Such an estimator is frequently used in the recent literature by many authors (Efron and Tibshirani, 1993, to mention a few). p_n , under H_0 has a distribution depending on α , while under H_1 it assumes various forms, dependent on α , $n \in F$; moreover, p_n tends to 1 to the limit on n . The different estimators can be more or less efficient, but the kernel of the question is not about these characteristics, but whether or not to use them.

1.14 The PHP is often requested to be added to the result $\Phi_\alpha=0$. Recently, the PHP was criticized by some authors, mainly for the following reasons:

- (i) PHP is used in a contorted way to calculate confidence intervals (Hoenig and Heisey, par. 2.1, 2001).
- (ii) It is thought that the information provided by the PHP was tautological in respect to the test results (Goodman and Berlin, 1994). In fact, for one-tailed tests, there is:

$$\Phi_\alpha(X_1, \dots, X_n) = \begin{cases} 0 & \Leftrightarrow p_n < 1/2 \\ 1 & \Leftrightarrow p_n \geq 1/2 \end{cases}$$

- (iii) It has been proven, for more simple formulation tests, (ex. Z-test) that there is a decreasing monotone function $g:(0,1) \rightarrow (0,1)$, with $g(0^+)=1$ and $g(1^-)=0$, such that $g(a_n)=p_n$ (Hoenig and Heisey, 2001.) This means that the information in a_n and p_n is the same.

- (iv) The suggestion of increasing the sample size in an opportune way, to then permits a significant test to be found in a future experiment, when a negative outcome $\Phi_\alpha=0$ is observed with a p-value close α (that, as a consequence of (iii), implies via (ii), to have a PHP in a one-tail test a little less than $1/2$), has been criticized.

1.15 Our response to the preceding points are the following:

- (i) We are completely in agreement with the criticism. Yet, an incorrect use of a method must in no way negate its validity when used correctly.
- (ii) It is true and we agree. Nevertheless the p-value is tautological with respect to the test results (see the definition 1.7). Also, some authors define the actual test departing from the p-value.
- (iii) We admit that there does exist such a function g for every statistical test. So, PHP does not add information to the p-value. Nevertheless, the point of view furnished by the PHP is different from that given by the p-value, in that the first is rational, and the second is emotional. (see points 1.9-12)
- (iv) We maintain that increasing the sample size in a successive experiment, calculating it in such a way as to obtain a pre-fixed (and estimated) power, would be absolutely correct. In fact, the sample size resulting from this method, is a consistent estimator of the unknown parameter, represented by sample size that one obtains when the distribution F is known (under the hypothesis $F \in P_I$).

Applying the plug-in estimator, one obtains strong consistency (De Martini, 2004). On the other hand, when H_0 is true, remember that the outcomes of the repeated statistical tests behave as a sequence of Bernoulli independent random variables with parameter $\alpha' \leq \alpha$ (see 1.3).

1.16 As a consequence of what have been said, considering the indicator of “stability” PHP in the case in which $\Phi_\alpha=1$, and evaluating to the contrary 1-PHP when $\Phi_\alpha=0$, one would have an estimate of the *reproducibility* of the statistical decision in a test analogous to the one just done. The reproducibility is a basic concept of an experimental result, making reference to the Galilean scientific method.

1.17 Furthermore, in our opinion, the PHP, in terms of reproducibility, would make researchers more *conscious*, clarifying that the outcome of a statistical test is only *one* observation of a random variable with two possible outcomes. This would bring, in our judgement, a more *mature* application of the statistical test tool.

1.18 The extension to k-samples tests is straightforward. In fact in that case, the plug-in estimator of the power results in being a generalized V-statistic, with the consequent properties of consistence (De Martini, 2004).

1.19 In Bayesian hypothesis testing, the statistical decision depends on the prior distribution on H_0 and H_1 , which defines A_n and R_n . Then, the test is no more dependent on α , as the power π_n does. As a consequence, instead of using the p-value, the Bayes factors are computed (Berger, 1985). But, from 1.10-12 and 1.16-18 it follows that the PHP should be considered as a stability index.

1.20 Under the computational profile, the calculation of the p-values falls on purpose to define the test results (according to the definition of test in 1.7). In fact, it is much more simple to implement the distribution of the test statistic (and thus calculate the p-value by integrating the test statistic distribution under H_0 over an interval limited by the observed test statistic) rather than to implement the definitions of the regions of Acceptance and Rejection. On the contrary, the power of the tests often has complex formulations, becoming undefinable in nonparametric contexts, being F unknown. Consequently, one can object that the proposal of computing PHP is null or has little value because it is non-generalizable. Rather, using the simple estimator based on the plug-in (see 1.13), one obtains a general and consistent solution. For the calculation of the PHP obtained from the plug-in, one can apply the Monte Carlo method, the approximation of which can be controlled in the classical way.

1.21 Finally, in classical hypothesis testing, the use of PHP cannot prescind the setting of the threshold α . This would minimize the temptation to “adjust” the type I error in function of the p-value observed.

1.22 In conclusion, to interpret the statistic test results, we are more likely to take a rational approach based on the calculation of the PHP. This is also because it is allowed by probability theory and modern calculation means.

References

- Berger J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer, NY.
Cohen J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NY.
Collings B.J., Hamilton M.A. (1988) Estimating the power of the Two Sample Wilcoxon Test for Location Shift, *Biometrics*, 44, 847-860.
De Martini D. (2004) A general method to estimate the power and the sample size of statistical tests, *Quaderno di Dipartimento SEMEQ*, n.75.
Efron B., Tibshirani R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, NY.
Lehmann E.L. (1997) *Testing Statistical Hypotheses*, Springer, New York .
Goodman S.N., Berlin J.A. (1994) The Use of Predicted Confidence Interval When Planning Experiments and the Misuse of Power when Interpreting Results, *Annals of Internal Medicine*, 121, 2, 200-206.
Hoenig J.M., Heisey D.M. (2001) The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis, *The American Statistician*, 55, 1, 19-24.
Ruhkin A.L. (1993) Estimation of the noncentrality parameter of an F-distribution, *Journal Statistical Planning and Inference*, 35, 201-211.
Yuan K.H., Maxwell S. (2004) On the Post-Hoc Power in testing Mean Differences, *Journal of Educational and Behavioral Statistics*, to appear.