STATISTICA, anno LXXVI, n. 4, 2016

THE USE OF P-VALUES IN APPLIED RESEARCH: INTERPRETATION AND NEW TRENDS

Donata Marasini
Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Università degli Studi di Milano-Bicocca, Milano, Italia

Piero Quatto
Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Università degli Studi di Milano-Bicocca, Milano, Italia

Enrico Ripamonti [1]
Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Università degli Studi di Milano-Bicocca, Milano, Italia

1. Introduction

In the early 2000s scholars started to discuss and criticize the use of the p-value in applied research in fields like Psychology, Ecology, and, more in general, in life sciences and experimental contexts. This criticism came from prestigious journals, such as *Epidemiology* in 2001, *Ecology* in 2014, and *Basic and Applied Social Psychology* in 2015. In particular, it was questioned the validity of the use of p-values (as well as significance testing and confidence interval procedures) in applied disciplines.

Also the American Statistical Association (ASA) took an official stance at this regard, and in 2016 it was published a statement as the Editorial of the journal *The American Statistician* (Wasserstein and Lazar, 2016). This statement is composed by a main text, and opinions and observations included as supplementary material.

As it will emerge in the following paragraphs, part of these discussions led to a reappraisal of the very notion of p-value, as originally proposed by Fisher starting from the 1920s, and subsequently by Neyman and E. Pearson.

This paper is organized as follows. In section 2 we will introduce and comment the ASA statements on the p-value. In section 3 we will consider possible alternatives to p-values, introducing the Bayes Factor, which is a classical instrument in Bayesian inference. In section 4 we will summarize a new approach recently put forward in the literature (Bayarri *et al.*, 2016) that seems particularly interesting in solving some of the problems related to the classical notion of p-value. Finally, section 5 is dedicated to draw some conclusions.

[1] Corresponding Author. E-mail: enrico.ripamonti@unimib.it

2. The ASA statement

The ASA statement is made up by 6 main points, summarizing the properties and shortcomings of the notion of p-value, which are also due to misinterpretations and misuses. Other misconceptions had been already put forward in the literature (e.g., Goodman (2008); Greenland *et al.* (2016)).

The ASA proposed a very general definition of p-value, in terms of "the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value", and put forth the following points and comments (reported below in italics, together with other comments).

1. P-values can indicate how incompatible the data are with a specified statistical model.

A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

P-value is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

In more formal terms, point 2 implies that if p is the probability that the test statistic D is higher or equal than d_0 (the observed value) when the null hypothesis H_0 is true, then p cannot directly indicate the probability associated with H_0.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect.

Indeed, the very notion of p < 0.05 does not imply that the null hypothesis is false, but that data are unusual in the light of the hypothesis and with the assumptions underlying this hypothesis, or that data are not consistent with these assumptions.

4. Proper inference requires full reporting and transparency.

Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherrypicking promising findings, *also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "p-hacking" leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided.*

In particular, the practice of p-hacking (Motulsky, 2015), which consists in analyzing and re-analyzing data with the sole purpose of obtaining a significant p-value, dramatically impoverishes the statistical analysis in favor of a mere research of significance, which is a well-established practice among scientific journals that conceive statistical significance as a necessary condition for the publication of a paper: "Obtaining a p-value that indicates that statistical significance is often requirements for publishing in a top journal" (Vidgen and Yasseri, 2016)

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise.

To provide an example with respect to the previous point, let's consider a trial in which a new treatment is compared with a placebo. This can be realized by observing the number of successes in a sequence of independent binomial trial for each treatment: $Y_0 \sim Bin(n_0, \omega)$, and $Y_1 \sim Bin(n_1, \omega + \theta)$. We consider the null hypothesis on the shifting parameter θ: $H_0 : \theta = \theta_0$ and the test statistic:

$$\left| \frac{Y_1}{n_1} - \frac{Y_0}{n_0} - \theta_0 \right|$$

measuring the distance between the difference of proportions and the corresponding difference of probabilities specified by H_0. For large sample sizes, a straightforward Normal approximation allows the tail probability above to be easily computed as

$$p(\theta_0) = 2 \left[1 - \Phi \left(\frac{|\frac{y_1}{n_1} - \frac{y_0}{n_0} - \theta_0|}{\sqrt{\frac{y_0(n_0-y_0)}{n_0^3} + \frac{y_1(n_1-y_1)}{n_1^3}}} \right) \right]$$

where Φ denotes the standard Normal cumulative distribution function, y_0 and y_1 are the observed values and $p(\theta_0)$ represents an asymptotic p-value for testing H_0. If the null hypothesis is $\theta = 0$, and the observed values are $\frac{y_0}{n_0} = 0.50$ and $\frac{y_1}{n_1} = 0.64$, with $n_0 = n_1 = 100$, the p-value is $p(0) = 0.043$. With the same null hypothesis, if $\frac{y_0}{n_0} = 0.50$ and $\frac{y_1}{n_1} = 0.5045$, with $n_0 = n_1 = 100,000$, we would still obtain $p(0) = 0.043$. However, if we use as a measure of effect the difference between proportions, it emerges a different result in the two different scenarios, which is not highlighted by the p-values. In addition, if we employ a classical effect size measure (Cohen, 1988):

$$h = |\varphi_1 - \varphi_2|$$

with $\varphi_i = 2\arcsin\sqrt{\hat{p}_i}$, being $\hat{p}_i$ $(i = 1, 2)$ the sample proportion, we would have $h_1 = 0.284$ in the first case, and $h_2 = 0.009$ in the second case.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.

For instance, if we consider a Normal random variable with location parameter θ, and we draw a random sample from this variable, obtaining a sample mean $\hat{x} = 2$, the θ value with maximum degree of compatibility with the observed value is in correspondence with the observed sample mean.

In the ASA statement, together with the previous six points referred to p-values, it is also mentioned a list of possible alternative procedures, i.e., "confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors and other approaches such as decision-theoretic modeling and false discovery rates". Other authors have recommended "strong descriptive statistics, including effect sizes" (Trafimow and Marks, 2015) or "practices, including estimation based on effect sizes, confidence intervals, and meta-analysis" (Cumming, 2013).

3. The Bayes factor

The proposal that will be presented in section 4 keeps into account both the frequentist p-value and the Bayesian approach. The Bayes Factor (B) represents indeed a compromise between the frequentist and the Bayesian perspectives (Goodman, 1999), although some authors have conceived the B in the light of the likelihood approach (Johnson, 2016) and other researchers as a measure of evidence in the frequentist field (Bayarri *et al.*, 2016). Accordingly, the B can be defined in a totally general manner as:

$$B = \frac{P(Data|H_0)}{P(Data|H_1)} \tag{1}$$

In case the null hypothesis and the alternative hypothesis are both composite, 1 is a likelihood ratio in which the numerator and the denominator are weighted by the respective prior. In case of a simple null hypothesis and composite alternative hypothesis, i.e.,

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0 \tag{2}$$

the B is:

$$B = \frac{L(x|\theta_0)}{m(x)} \tag{3}$$

where

$$m(x) = \int_{\theta \neq \theta_0} L(x|\theta)\pi(\theta)d\theta \tag{4}$$

and $\pi(\theta)$ indicates the prior distribution with respect to H_1. In case both the hypotheses are simple, 3 is a likelihood ratio. In case the distribution under study is Normal with known variance, under hypotheses 2 it can be shown that the minimum of 3 is:

$$B = e^{-1/2z_{p/2}^2} \tag{5}$$

where $z_{p/2}$ is the order $p/2$ quantile of the standard Normal (Goodman, 1999). Moreover, if the p-value is *proper* (Bayarri *et al.*, 2016), i.e.,:

$$P(p \leq \alpha|H_0) = \alpha$$

to the p-value is associated, under the null hypothesis, a continuous uniform distribution function, defined in the interval $(0, 1)$, so that the null hypothesis in 2 can be substituted with $H_0 : \varphi(p) = 1$ $(0 < p < 1)$. As to the alternative hypothesis in 2, it has been suggested to set a Beta distribution with parameter $(\omega, 1)$ so that H_1 can be written as $H_1 : \varphi(p) = \omega p^{\omega-1}$ $(0 < p < 1)$. Such Beta distribution is decreasing in p in the $(0, 1)$ interval, hence to high values of the test statistic are associated low values of p under the alternative. So, 2 can be written as:

$$H_0 : p \sim U(0, 1) \quad H_1 : p \sim \text{Beta}(\omega, 1) \quad (0 < \omega < 1)$$

The parameter ω can be modeled with a prior $\pi^{'}(\omega)$, and 3 can be written as:

$$B = \frac{\varphi(p|H_0)}{\int_0^1 \varphi(p|H_1)\pi'(\omega)d\omega} = \frac{1}{\int_0^1 \omega p^{\omega-1}\pi'(\omega)d\omega} \tag{6}$$

It has been shown (Sellke *et al.*, 2001) that 6 takes its minimum at $-\mathrm{e}p\log p$ for $p < \frac{1}{\mathrm{e}}$ and for every $\pi^{'}(\omega)$. In the light of 1 and 3 it follows:

$$B \geq -\mathrm{e}p\log p \tag{7}$$

It is very important to consider inequality 7 for inferential purposes: in fact, in case there was not a suitable prior $\pi(\theta)$ to calculate the B, at least a minimum bound can be given. In case this minimum is very low, this should lead to consider the alternative hypothesis instead of the null hypothesis (the so-called "surprise effect", Bayarri and Berger (1999)). It is worth observing that the minimum 7 is attained in case of bidirectional alternative hypothesis, the reader can refer to Benjamin and Berger (2016) for a discussion of the unidirectional case.

It has been also suggested (Colquhoun, 2014) to define 1 as:

$$B = \frac{P(p \leq \alpha|H_0)}{P(p \leq \alpha|H_1)} \tag{8}$$

with

$$P(p \leq \alpha|H_1) = 1 - \bar{\beta} = \int [1 - \beta(\theta)]\pi(\theta)d\theta \tag{9}$$

TABLE 1
Values of B in the light of different p-values (Goodman, 2001).

p-value	$B = e^{-1/2 z^2_{p/2}}$	$B = -ep \log p$
0.1	0.26	0.62
0.05	0.15	0.406
0.01	0.036	0.125
0.001	0.005	0.016

and, in case of a proper p-value, 8 can be written as:

$$B = \frac{\alpha}{1 - \bar{\beta}} \tag{10}$$

As we have briefly summarized, the B is not an univocal concept in statistical theory. Applying definition 3, for instance in case of $B = 0.5$, one could conclude that data provide a double support to the alternative with respect to the null hypothesis.

In case of a Normal distribution, defining the B by 5 with $p = 0.05$ and $z_{p/2} = 1.96$, it would follow that $B = 0.15$, a value that has been assessed as "moderate strength of evidence" (Goodman, 1999). In Table 1 we show the minimum value for expression 5 and 7 for different p-values. Relevant differences in terms of minimum point did emerge; for instance, when $p = 0.05$, data support the null hypothesis about three times more in case $B = -ep \log p$ than assuming a Normal distribution. Goodman (2001) does not take a definitive position on which minimum to adopt, however, as it will emerge, to calculate minimum 7 it is important to consider some precautions.

4. A new proposal

Bayarri *et al.* (2016) recently put forth a new proposal which is worth reviewing; in a very challenging paper the authors posit a "simple modification of standard methods". In a first stage, the researcher should explain the experimental design to follow in his/her research. At this regard, the authors introduced the "pre-experimental" rejection odds (O_{pre}), which is given by the product of the prior odds for the alternative hypothesis (π_1) with respect to the null hypothesis (π_o), i.e., $\frac{\pi_1}{\pi_0} = \frac{1-\pi_0}{\pi_0}$ and a "pre-experimental" rejection rate (R_{pre}), given by the ratio $\frac{1-\bar{\beta}}{\alpha}$, reciprocal of 10. Thus,

$$O_{pre} = \frac{\pi_1}{\pi_0} \times \frac{1-\bar{\beta}}{\alpha} = \frac{\pi_1}{\pi_0} \times R_{pre} \tag{11}$$

In 11, in the authors' terms, it is indeed identified the "odds of correct rejection of the null hypothesis to incorrect rejection". Thus, fixing for instance $\alpha = 0.05$ and $1-\bar{\beta} = 0.80$ and if the odds of the alternative compared to the null is $\frac{\pi_1}{\pi_0} = 1$, i.e., if the two *a priori* assumptions are equal, then $O_{pre} = 16$: this value can be interpreted stating that the correct rejection of the null hypothesis is 16 times higher with respect to the incorrect rejection. If the odds is $\frac{\pi_1}{\pi_0} = 0.0625$, i.e., *a priori* the probability associated to the null hypothesis is 0.941, with power again $1 - \bar{\beta} = 0.80$, it would follow $O_{pre} = 1$, which means that the correct rejection and the incorrect rejection of the null hypothesis have the same probability, hence giving to the alternative hypothesis a lower support than the previous case.

It is worth observing that O_{pre} is extremely sensitive to the values of all the parameters involved, i.e., π_1, α, $1-\bar{\beta}$. At this regard, let's consider the two following examples. If, in the previous case (with $\alpha = 0.05$ and $\frac{\pi_1}{\pi_0} = 1$) the power were 0.2, it would follow $O_{pre} = 4$, hence the correct rejection rate would lower considerably. It is not so infrequent to have a power value of 0.2, as reported by the following authors: "in practice, many published results have a power far less than 0.8. Values around 0.5 are common, and 0.2 is far from rare" (Colquhoun, 2014); "we optimistically estimate the median statistical power of studies in the neuroscience field to be between about 8% and about 31%" (Button *et al.*, 2013). In many research fields, such as experimental physics or genomics, it is a common practice to give much support to the null hypothesis and set the probability level α very low. For instance, following a case study reported in Bayarri *et al.* (2016) in the field of genomics, the ratio $\frac{\pi_1}{\pi_0}$ was 10^{-5}, hence involving an "almost" certainty on the null hypothesis, the power was $1-\bar{\beta} = 0.5$ and $\alpha = 5 \times 10^{-7}$ with an *a priori* support for H_0: $O_{pre} = 10$. We underline that this value would become $O_{pre} = 100$ if we fix $\alpha = 5 \times 10^{-8}$, as in several genomics applications. It is clear that the ratio 100:1 is a solid guarantee to refer, at least *a priori*, to a new discovery.

We remark that, even though the ratio R_{pre} does coincide with a B, it is employed only as a tool for *a priori* comparisons of rejection (correct or incorrect) of the null hypothesis.

Finally, the ratio R_{pre} could be actually employed also in a frequentist scenario, as suggested by the authors. In these terms, the prior $\pi(\theta)$ related to $1-\bar{\beta}$ should be fixed to a single value θ_1 belonging to the space Θ_1. In this case, given $R_{pre} = \frac{1-\bar{\beta}}{\alpha}$ it would be worth fixing the two probabilities so that $R_{pre} > 1$, hence leading to a correct test (in frequentist sense) for the hypothesis $H_0 : \theta = \theta_0$ vs. $H_1' : \theta = \theta_1 \neq \theta_0$.

After the "pre-experimental" phase of research, it would follow a "post-experimental" phase in which it is introduced a "post-experimental" rejection odds (O_{post}), which is given by:

$$O_{post} = \frac{\pi_1}{\pi_0} \times R_{post} \tag{12}$$

where R_{post} indicates the B in the form 3 of H_1 with respect to H_0.

The authors report the following example which is worth considering. Let's consider a trial involving a new HIV vaccine; to a first group of 8,197 individuals (51 of whom HIV-positive) it is prescribed the new vaccine, whereas to a second group of 8,198 individuals (71 of whom HIV-positive) it is prescribed a placebo. We use the Normal approximation to test the hypothesis of absence of effect ($H_0 : \theta = 0$) versus the unidirectional hypothesis indicating presence of the effect ($H_1 : \theta > 0$). Fixing *a priori* $\alpha = 0.05$ and $\frac{\pi_1}{\pi_0} = 1$, it is obtained $1-\bar{\beta} = 0.45$. It follows $O_{pre} = 9$, i.e., the odds for a correct rejection is nine times higher than an incorrect reject. After the trial, the authors obtained a z-value $z = 2.06$, which implies a p-value of 0.02. The authors also propose three different values for R_{post}, which are related to the prior $\pi(\theta)$ and that can be set in three different ways: (i) using an empirical-based prior, it would

TABLE 2
Values of R_{post} with different alternative hypotheses.

H_1	R_{post}
$\theta = 2$	0.03
$\theta = 3$	49.40
$\theta = 4$	24.29
$\theta = 5$	4.45

follow $R_{post} = 4$; (ii) using a uniform prior in the interval $(0, 2.95)$, it would follow $R_{post} = 5.63$; (iii) using a prior $\pi(\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimation, it would follow $R_{post} \leq 8.35$. Since $\frac{\pi_1}{\pi_0} = 1$, applying 12 the three possible values of O_{post} are given by: $O_{post} = 4$, $O_{post} = 5.63$, $O_{post} \leq 8.35$, thus disconfirming the value of $O_{pre} = 9$, obtained in the "pre-experimental" phase of research. Using the authors' words: "the pre-experimental 9 does not accurately represent what the data says". As to the maximum obtained as the reciprocal of 5 we remark that the observed value can be far from that assumed by the alternative hypothesis. Let's consider a Normal random variable of parameter $(\theta, 1)$ and the following hypotheses: $H_0 : \theta = 0$ vs. $H_1 : \theta = 1$. Imagine that a sample with $n = 1$ provides the value $x = 2.8$ thus leading to $R_{post} = 9.97$. In case of different alternative hypotheses, this value would also vary, as shown in Table 2.

The maximum value $R_{post} = 50.40$ would be achieved for $\theta = 2.8$ and values of the alternative hypothesis similar to the observed value lead to higher values of R_{post}.

Bayarri *et al.* (2016) also put forward a frequentist view of this approach, which can be obtained using 7, from which it follows:

$$R_{post} \leq \frac{1}{-\mathrm{e}p \log p} \tag{13}$$

and in this way R_{post} would depend upon the data only through the p-value.

For example, considering the values $\frac{\pi_1}{\pi_0} = 1$, $\alpha = 0.05$ and $1 - \bar{\beta} = 0.8$ fixed *a priori* and the p-value $p = 0.01$ obtained after the realization of the experiment, using 13 it would follow $O_{post} \leq 8.01$, which is very different from $O_{pre} = 16$. However, in case a researcher had obtained a p-value $p = 0.001$, from 13 it would have followed $R_{post} \leq 53.42$. Actually this inequality is not informative on the real value of R_{post} and this is a real challenge for this approach.

We remark that inequality 13 should be applied only when the conditions required by the inequality are satisfied. In fact, if we apply 13 to the previous example with $x = 2.8$ it would follow $p = 0.0052$ and a maximum value of 13.49. However, as R_{post} is given by:

$$R_{post} = e^{-1/2\theta^2 + 2.8\theta}$$

at least the hypotheses comprised in the interval $1.18 \leq \theta \leq 4.43$ lead to maximum values higher than 13.49.

5. Conclusion

In this paper we have summarized a recent controversy involving the concept of p-value and its use in applied research. From the one side, several authors recommended to adopt the p-value, but along with other procedures: "p-values, confidence intervals, and information theoretic criteria are just different ways of summarizing the same statistical information" (Murtaugh, 2014); "the p-value is a very valuable tool, but when possible it should be complemented - not replaced - by confidence intervals and effect size estimates" (Benjamini, 2016). It is also worth reporting the comment by Greenland *et al.* (2016): "we have no doubt that founders of modern statistical testing would be horrified by commons treatments of their invention", which highlights the misuses of p-value in applied research. From the other side, several authors criticized the use of p-values in applied research. For instance, Cumming (2013) asserted to "not trust any p-value" and "whenever possible, avoid using or statistical significance or p-value"; Trafimow and Marks (2016) claimed: "we reiterate the message from our 2015 editorial. The ban of p values continue".

In the second part of this paper we reviewed the proposal by Bayarri *et al.* (2016), which seems very modern and promising. Indeed, this proposal adopts instruments already known in the statistical literature and does "not require any changes in the statistical tests that are commonly used, and would rely only on the most basic statistical concepts and tools, such as significance thresholds, p-values, and statistical power".

As we have seen, the authors introduced O_{pre}, which is an *a priori* odds (see formula 11) and O_{post}, which is an *a posteriori* odds (see formula 12). The innovation proposed by the authors is that of considering the BFin 8 in the version 11, where the values are to be set in the pre-experimental phase. Thus, the effect obtained after the realization of the experiment can be verified by comparing R_{pre} with R_{post}. Actually the framework of this proposal is Bayesian, but, as we described, the authors also posit a frequentist interpretation.

It is worth observing that criticism on this proposal has been put forth as to the assumptions that are necessary to attain the maximum 13, which have been judged as "entirely implausible" (Ioannidis, 2014). In our opinion such criticism is too extreme, but it is true that the proposal by Bayarri *et al.* should be further tested on the field, so that one can verify the real possibility of applying this framework to empirical problems.

In conclusion, in this article we revised some recent controversy on the use of p-values in applied research, starting from the analysis of papers published in *Ecology*, *Epidemiology*, and *Basic and Applied Social Psychology.* Despite of the criticism, p-values continue to be largely employed in applied research, as observed by Ioannidis (2014): "p-values continue to be widely used and misused, but until now there has been a lack of consensus in the scientific community. Many competing options exist to change the paradigm... The current status quo is perpetuated". Even though alternative and valid instruments have been proposed in the statistical literature, such as those that we reviewed in this paper, the shift from the "p-value paradigm" to other paradigms seems very slow. In our opinion the reason from this ineptitude to change has been effectively described by Goodman (2016): "Exactly how our scientists are supposed to do that? Where are all the textbook examples? Where are the examples in the published literature?"

References

M. Bayarri, D. J. Benjamin, J. O. Berger, T. M. Sellke (2016). *Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses.* Journal of Mathematical Psychology, 72, pp. 90–103.

M. Bayarri, J. O. Berger (1999). *Quantifying surprise in the data and model verification.* Bayesian Statistics, 6, pp. 53–82.

D. Benjamin, J. Berger (2016). *A simple alternative to p-values.* The American Statistician, Online supplement, 70.

Y. Benjamini (2016). *It's not the p-values fault.* The American Statistician, Online supplement, 70.

K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, M. R. Munafò (2013). *Power failure: why small sample size undermines the reliability of neuroscience.* Nature Reviews Neuroscience, 14, pp. 365–376.

J. Cohen (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

D. Colquhoun (2014). *An investigation of the false discovery rate and the misinterpretation of p-values.* Royal Society Open Science, 1, pp. 1–16.

G. Cumming (2013). *The new statistics why and how.* Psychological Science, 25, pp. 7–29.

S. Goodman (2008). *A dirty dozen: twelve p-value misconceptions.* Seminars in Hematology, 45, pp. 135–140.

S. Goodman (2016). *The next questions: Who, what, when, where and why?* The American Statistician, Online supplement, 70.

S. N. Goodman (1999). *Toward evidence-based medical statistics. 2: The Bayes factor.* Annals of Internal Medicine, 130, pp. 1005–1013.

S. N. Goodman (2001). *Of p-values and bayes: a modest proposal.* Epidemiology, 12, pp. 295–297.

S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, D. G. Altman (2016). *Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations.* European Journal of Epidemiology, 31, pp. 1–14.

J. P. Ioannidis (2014). *Discussion: why "an estimate of the science-wise false discovery rate and application to the top medical literature" is false.* Biostatistics, 15, pp. 28–36.

V. Johnson (2016). *Comments on the ASA statement on statistical significance and p-values and marginally significant p-values.* The American Statistician, Online supplement, 70.

H. J. Motulsky (2015). *Common misconceptions about data analysis and statistics.* British Journal of Pharmacology, 172, pp. 2126–2132.

P. A. Murtaugh (2014). *In defense of p-values.* Ecology, 95, pp. 611–617.

T. Sellke, M. Bayarri, J. O. Berger (2001). *Calibration of ρ values for testing precise null hypotheses.* The American Statistician, 55, pp. 62–71.

D. Trafimow, M. Marks (2015). *Editorial.* Basic and Applied Social Psychology, 37, pp. 1–2.

D. Trafimow, M. Marks (2016). *Editorial.* Basic and Applied Social Psychology, 38, pp. 1–2.

B. Vidgen, T. Yasseri (2016). *P-values: misunderstood and misused.* pre-print Cornell University.

R. L. Wasserstein, N. A. Lazar (2016). *The ASA's statement on p-values: context, process, and purpose.* The American Statistician, 70, pp. 129–133.

Summary

In this paper we consider a controversy on the use and interpretation of p-values in applied research. In recent years several applied and theoretical journals have started to discuss on the appropriate use of p-values in research fields such as Psychology, Ecology, and Medicine. First, the notion of p-value has some intrinsic limitations, which have been already highlighted in the statistical literature, but are far from being recognized in applied research. Second, it has emerged the so-called practice of p-hacking, which consists in analyzing and re-analyzing data until obtaining a significant result in terms of a p-value less than 0.05. In the light of these problems, we review two alternative theoretical frameworks, given by the use of Bayes factor and a recent proposal that leads to evaluate statistical hypotheses in terms of *a priori* and *a posteriori* odds ratios.

Keywords: p-value; Neyman-Pearson; Bayes factor; odds ratio; p-hacking.