ISSN: 2574-1241

DOI: 10.26717/BJSTR.2018.06.001359

Bruno Mario Cesana. Biomed J Sci & Tech Res

BIOMEDICAL
Journal of Scientific & Technical Research

Review Article Open Access

What p-value must be used as the Statistical Significance Threshold? P<0.005, P<0.01, P<0.05 or no value at all?

Bruno Mario Cesana*[1,2]

[1]*Former: Department of Molecular and Transactional Medicine, Statistics and Biomathematics Unit, Faculty of Medicine and Surgery, University of Brescia, Brescia, Italy*

[2]*Contract Professor: Department of Clinical Sciences and Community Health, Unit of Medical Statistics, Biometry and Bioinformatics "Giulio A. Maccacaro", Faculty of Medicine and Surgery, University of Milan, Milan, Italy*

Received: June 20, 2018; **Published:** July 06, 2018

***Corresponding author:** Bruno Mario Cesana, Department of Molecular and Transactional Medicine, Statistics and Biomathematics Unit, Faculty of Medicine and Surgery, University of Brescia, Brescia, Italy

Abstract

Statistical significance is a tool for making decisions on a probabilistic basis extensively used in the scientific world. It must be recognized that the 0.05 value as the threshold of the statistical significance is undoubtedly arbitrary and nothing prevents it from being modified according to well-founded arguments. Furthermore, it must also be recognized that the logic of the statistical significance test is quite debatable, as well as being little understood by researchers who are the main users. Also the meaning of p-values is often ignored with consequent misinterpretations and misunderstandings. I will give a general overview and some insights on the topics of the p-values and of the statistical significance.

Keywords: Statistical Significance; True Null Hypothesis; Bayes Factor; Prior and Posterior Probabilities; Prior and Posterior Odds; False Discovery Rate; Claims of New Discoveries; Irreconcilability; Null Hypothesis Significance Testing Procedure; Strength of Evidence

Introduction

Recently, seventy-two eminent biostatisticians, psychologists, philosophers, science methodologists, economist, etc. (let's define them, generally, as scientists) propose "to change the default P-value threshold for statistical significance from 0.05 to 0.005." [1]. The proposal has been motivated by the "the lack of reproducibility of scientific studies" and by the fact that "the statistically significance threshold of the P<0.05 gives a high false discovery rate", even in the absence of any flaws in the experimental design, conduction of the study, statistical analysis and reporting of the results. Moreover, it has to be stressed, that the proposal applies to "claims of new discoveries" and that "should not be used to reject publications of novel findings with 0.005<P<0.05 properly labeled as suggestive evidence". As a further remark, the expression "false discovery rate" misuses, as it happens very often, the term "rate" which appropriately means "a measure of the frequency per unit time of some phenomenon of interest" [2]; indeed, the appropriate term in this context is proportion. This paper [1] has been also commented by Ioannidis [3], who is also one of the authors of the paper [1], by re-iterating the criticisms of the results of the biomedical research expressed some years ago in a seminal paper titled "Most Published Research Findings Are False" [4]. However, even if this paper has been and it continues to be a very relevant landmark for this topic and debate, it seems to be not free from some methodological problems, particularly regarding the model employed for calculating the posterior probability, as Goodman and Greenland pointed out [5,6]. Moreover, it must also be said that some criticisms on the statistical testing paradigm, on the use of the statistical significance tests instead of the confidence intervals and on the abuse and misinterpretation of p-values have been raised for a long time.

To deepen these topics very useful references, among others, are the landmark paper by Berger and Sellke [7] with the very impressive title of "The Irreconcilability of P-Values and Evidence" together with six companion commentaries in the issue of March, 1987 of the Journal of the American Association and the paper by Goodman [8]. In addition, also the paper by Moran and Salomon [9] has to be particularly recommended, owing to its pretty exhaustive review of the statistical test theory also from the historical point of view. Indeed, the debate on the statistical significance test goes back to the first work of Fisher and to those of Neyman and Pearson together with the disputes between Fisher and Neyman (mainly), as it has been reported, among others, in some books [10-12] to which readers are referred. In addition, it has to be said that the proposal of moving the significance threshold from 0.05 to 0.005 has been previously formulated by Johnson [13], one of the authors of the paper of Benjamin et al. [1] This proposal has been commented by two letters [14,15] with a reply by Johnson [16] to which the interested readers are referred.

Finally, the very impressive Ioannidis's paper "Most Published Research Findings Are False" [4] has been commented by Jager & Leek [17] who reported a substantial reduction of the "false discovery rate" to 14% leading to the conclusion that "the medical literature remains a reliable record of scientific progress". However, the Jager and Leek's paper has been furtherly criticized by six companion commentaries [18-23] in the same issue of Biostatistics with, not surprisingly, very different judgements and considerations. Indeed, it is very well instructive to see how many aspects can be raised by a statistical method together with its practical realization. However, as a general conclusion, it seems that the drastic and dramatically alarming Ioannidis's statement [4] has to be mitigated to some extent. Coming back to the meaning and the interpretation of the p-values, it is important to stress that Ioannidis reported [3], according to Wasserstein & Lazar [24], that the most common misinterpretation of p-values, among the multiple ones present in the scientific literature, is that they represent the "probability that the studied hypothesis is true".

So, according to this misunderstanding, "a P value of .02 (2%) is wrongly considered to mean that the null hypothesis (eg, the drug is as effective as placebo) is 2% likely to be true and the alternative (eg, the drug is more effective than placebo) is 98% likely to be correct" [2]. These wrong interpretations are not surprising since it is very well known the poor feeling that researchers have for Statistics and for the scientific reasoning based on the statistical methodology. Also some comments, particularly raised in the case of negative controlled clinical trials and only based on some clinical reasoning [25] without considering the corresponding statistical aspects [26], turn out to be rather questionable or, at least, definitely incomplete.

The difficulties of a correct interpretation of the p-values even led to banish the p-values from the Basic and Applied Social Psychology (BASP) journal; indeed, after a grace period of one year, announced by the first Trafimow's Editorial [27], the editors announced that BASP "would no longer publish papers containing P-values, because the values were too often used to support lower-quality research" [28]. Furthermore, in their Editorial, the Editors emphasized that "the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it." So, if this decision will be shared by other journals, we can also arrive at a situation of no p-values at all in the papers of the scientific literature. The BASP journal announcement has been commented by Nature [29]. This fact confirms the attention of Nature to the role of the Statistics in the scientific research and to the meaning of the p-values, as the publication of the paper by Nuzzo [30] and of the companion editorial [31] furtherly attests.

Nuzzo's paper [30] succeeded in drawing the attention of a large audience of physicians on the Bayes's rule, previously introduced in the epidemiological context by some papers written by Goodman [32-35]. In fact, Nuzzo's paper [30] shows very clearly, in a figure, how p-values of 0.05 or 0.01, empirically obtained from a statistical analysis, can modify three values of the a "priori odds" that the null hypothesis (H_0) is true; namely: "19 to-1 odds against the null hypothesis", "1-to-1 odds", and, finally, "9-to-1 odds in favor of the null hypothesis". I think that discovering that a p-value of 0.05 or even 0.01 can have a very little impact on the plausibility of an almost unlikely null hypothesis (19 to-1 odds or P=0.95 against) and that only in the case of a very plausible H_0 (9-to-1 odds or P=0.90 in favor) the p-values are very similar to the H_0 probabilities, could have made it clear the difference between the statistical significance and the probability that H_0 has of being true. Furthermore, Nuzzo's paper [30] made it clear that:

a) the statistical test is carried out considering true the null hypothesis

b) how this assumption is in fact questionable

c) Finally, how it is practically not sensible to reason in terms of a "true null hypothesis" for concluding in the terms of the evidence of a clinical research. Indeed, the message very well spread by this paper is that the null hypothesis, assumed to be absolutely true under the paradigm of the statistical significance test, has actually an unknown probability of occurring and that it is sensible to consider different probability scenarios of the veracity of H_0.

The only criticism that could be done on Nuzzo's paper [30] consists in the fact that have not been shown the formulas of the Bayes factor, leaving its role not very well defined; in addition, it has not reported which Bayes factor has been used for the calculations shown in the figure. Indeed, the pertinent answers to the questions related to the statistical methodology must be found by the reader in the referenced papers. An additional merit of Nuzzo's paper [30] was of leading the American Statistical Association to express its official position and thought about the meaning of the p-values in some papers [24,36] and also to publish on YouTube a very instructive video of the statistical section "ASA statement on P-values and statistical significance: Development and impact" with speakers Nuzzo, Johnson, and Senn [37].

A further explanation of p-value has been given by Mark et al. [38] and also a non-technical introduction to the p-value statistics has been reported by Figueiredo Filho et al. [39]. In addition, several formally correct videos on the topic of the p-values are on YouTube [40-42] together with one very amusing featuring cartoons as protagonists [43]. I do not want to make considerations about the philosophy of the science or on the role of the Statistics in the scientific research or to propone a new paradigm of the scientific method. Furtherly, I must say that I do not even share the controversy raised by some statisticians who would like only the intervals of confidence to be used instead of the statistical tests, because I think that both must be used, given that both provide useful information about the results of the statistical analysis of a research. Indeed, the problem is always of interpreting correctly the results of the statistical procedures and of knowing their meaning. As a biostatistician, more oriented in sample size calculations and clinical trials methodology, my aim is to point out the correct interpretation of the p-values together with some personal suggestions about their use focused also on the plausibility of the null hypothesis or to the probability that a null hypothesis has to be true.

P-values: Some Historical Considerations

According to Fisher [44] the p-values could be considered as an index of the "strength of the evidence" against H_0. Particularly, after having choose the statistical test, carried out the experiment, calculated the test statistic from the actual experimental data and the probability value associated with the test statistic, if this probability value is quite small (say, ≤0.05) the null hypothesis could be rejected. However, it would be better to use the expression "to not accept", according to a less strong expression that is more relevant to the probabilistic nature of the statistical testing procedure. Actually, Fisher popularized the use of the p-values in statistics and, particularly in his influential book Statistical Methods for Research Workers [45], proposed the level p = 0.05, or a "1 in 20 chance of being exceeded by chance", as a limit for statistical significance.

Then Fisher reiterated the p = 0.05 threshold and explained its rationale, stating: "It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results" [44]. So, a p-value of ≤0.05 on the null hypothesis indicated, according to Fisher [44], that: "Either an exceptionally rare chance has occurred or the theory is not true". Fisher's further advice [44] was that "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty…or one in a hundred".

Furtherly, in the Statistical tables for biological, agricultural and medical research compiled with Yates [46] there are reported the quantiles of several probability distributions (standardized Gaussian: (Table II). The Normal Probability Integral; Student's t: (Table III). Distribution of t; χ^2: (Table IV). Distribution of c2, and F: (Table V). Distribution of z and Variance Ratio for 20%, 10%, 5%, 1%, and 0.1%; thereafter it was called F distribution in honor of Fisher or F distribution shortly for the distribution of Fisher and Snedecor) for selected probability values. So, the computed values of the statistical tests could be compared against some cut-offs corresponding, especially, to the p-values of 0.05 (mainly) and 0.01, cementing their use as statistical significance thresholds.

A basic point, perhaps not very well understood, is that the inference from the p-value involves only the null hypothesis and that the "likelihood" of this hypothesis, calculated from the experimental data, is not also the "probability of the null hypothesis of being true". That is, the p-values should not to be misinterpreted as posterior probabilities that have to be obtained according to the Bayesian paradigm. However the main relevant and frequent use of the p-values is currently in the context of the Neyman-Pearson hypothesis testing frequentist paradigm, in which two hypotheses are formalized; namely the null hypothesis (H_0) and the alternative (H_1 or H_A), with the first to be tested versus the latter. Then, the test statistic is obtained from the formula of the pertinent statistical test and the corresponding probability value is calculated by referring to the probability distribution of the test statistic. It has to be pointed out that currently, the p-values are compared with the prefixed significance level instead of comparing the test statistics with the tabulated critical values that delimit the critical region of rejection (not acceptance) of H_0 of the pertinent probability distribution.

In fact, the diffusion of statistical software that calculates the probability values has made the consultation of tables quite obsolete. Furthermore, one thing is to state that the p-value is <0.05 and another is to report its exact value (to a certain number of decimal places) such as p = 0.0253. It has to remember that the critical region corresponds to an area of a probability distribution, and, therefore, to a probability value that is equal to the significance level, chosen by the researcher and defined as α at the left or right tail of the distribution in the case of a unilateral test or equal to α/2 at the left and right tail of the distribution in the case of a bilateral test. In the frequentist paradigm, are relevant the type I error (α) that corresponds to the probability of rejecting (not accepting) a true null hypothesis, and the type II error (β) that corresponds to the probability of not rejecting a false null hypothesis or, more known and quoted, the power of the statistical test given by 1-β. In fact, this procedure refers to the repetition of the same experiment carried out under the same conditions on samples repeatedly and randomly obtained from the same distribution (H_0 is true) or from two (at least) different distributions just in agreement to the alternative hypothesis.

Finally, it has also to consider the Jeffreys's approach to testing in the Bayesian context [7]. This method requires the definition of the Bayes factor as the ratio between the value of the maximum likelihood calculated from the experimental data under the null hypothesis (given the parameter under H_0 equal to θ_0, say) and the value of the likelihood calculated from the experimental data under the alternative hypothesis (given the parameter under H_1 equal to θ_1, say). Then the null hypothesis is rejected if the calculated ratio is <1 or, otherwise, if the value of the calculated ratio is >1, the null hypothesis is not rejected. Thereafter, it is possible to report the posterior probabilities (the probability that H_0 is true given the experimental data) by transforming the odds (the calculated Bayes factor) in a probability by recalling that probability (p) is obtained as p = 1 / (1 + odds). Finally, is also possible to calculate the posterior probability for the alternative hypothesis of being true, given the experimental data.

Considerations about the disagreement and the sparse points of agreement among the three giants Fisher, Neyman (with also E. Pearson) and Jeffreys are out of the limits of this editorial. Useful papers for some further insights are from Berger and Sellke [7], Hubbard and Bayarri et al. [47], Gibbons[48], Pratt et al. [49], De Groot [50], Christensen [51] and finally, Berger [52]. It has to be stressed that instead of *to reject*, I always used the expression *to not accept* just for underlying the probabilistic nature of the statistical testing procedure; otherwise, it has to be used the expression *not rejected* since the expression *to accept* has to be absolutely avoided owing to the fact that, according to the scientific paradigm, the null hypothesis can be only disproved. However, it has to be said that the expression *accept the null hypothesis* is currently used also in the statistical literature [53]. Finally, it has to be point out the fact that if the null hypothesis is not rejected, nothing could be concluded, and this is a point not well understood by clinical researchers. To this point it has to be remembered the sharp and definite sentence: "the absence of the evidence is not the evidence of the absence".

The meaning of the p-values

The p-value quantifies the probability of having obtained the experimental results "under the null hypothesis (H_0)" that is, usually, a hypothesis of no difference. Let's disregard for sake of simplicity the case of the non-inferiority settings in which the null hypothesis is of the "maximal difference not clinically/biologically relevant" and the recently considered superiority statistical testing in which the null hypothesis is of the "minimal difference clinically/biologically relevant" [54]. The expression "under the null hypothesis (H_0)" can be better paraphrased as "if the null hypothesis is true" but it is, maybe, only with the expression "given that the null hypothesis is true (p-value | H_0) that it is very well stated and understandable that the p-value of the statistical test is obtained as a conditional probability.

So, considering the formula of the conditional probability, the probability of H_0 of being true is equal to 1 [$P(H_0) = 1$] by the assumption underlying the statistical test of significance, and, consequently, the probability of the joint event given by a statistical significant result [$P(x_{OBS}) \leq 0.05$)] and H_0 true [$P(H_0) = 1$], defined as [$P(x_{OBS} \cap H_0)$] remains equal to the simple probability value (p-value) associated with the test statistic. Being the formula of a conditional probability given by:

$$P(x_{OBS} \mid H_0) = \frac{P(x_{OBS} \cap H_0)}{P(H_0)} \qquad (1)$$

Where x_{OBS} indicates the observed Result. It is well evident that a value equal to 1 at the denominator does not change the value at the numerator. It is also evident that the assumption "the null hypothesis is true" is useful for carrying out the test of significance and for being able to conclude against the null hypothesis or to make no conclusions at all. However, it is also well evident that in the real world there cannot exist a "true null hypothesis" or a "true alternative hypothesis". It is possible to argue that there is a situation "equipoise-like" in which the two hypotheses are equally probable of being true [$P(H_0) = P(H_A) = 0.5$] or situations in which $P(H_0) > P(H_A)$ or $P(H_0) < P(H_A)$, taking also into account the context of the research. So, it must reasonably be said that this paradigm is a useful tool for concluding about a research (a decisional rule on a probabilistic basis) but it is not adequate to conclude on the veracity of the null hypothesis. To this aim it has to consider a different approach built on the Bayesian theory.

Bayes Factors

For calculating the probability of the null hypothesis of being true it is necessary to refer to the "Bayes factor" that represents the evidence from the data, and the value of the "prior odds" that has to be obtained, according to Benjamin et al. [1] "by researchers' beliefs, scientific consensus, and validated evidence from similar research questions in the same field." Benjamin et al. [1] shows the application of the calculations focused on the truth of the alternative hypothesis (H_1) against the null hypothesis (H_0), but for keeping consistency with the familiar statistical testing paradigm focused on the null hypothesis, I will consider the opposite situation of the truth of the null hypothesis (H_0) against the alternative hypothesis (H_1), which involves the reversal of the likelihood ratio. So:

$$\frac{P(H_0 \mid x_{OBS})}{P(H_1 \mid x_{OBS})} = \frac{P(H_0)}{P(H_1)} \frac{f(x_{OBS} \mid H_0)}{f(x_{OBS} \mid H_1)} = BF(x_{OBS}) \bullet \text{prior} \quad \text{odds}(H_0)$$
$$\text{where} \quad \frac{f(x_{OBS} \mid H_0)}{f(x_{OBS} \mid H_1)} = BF(x_{OBS}) \rightarrow \text{Bayes Factor} \qquad (2)$$

Where the Bayes factor has to be calculated by considering the distributional properties of the observed data. It has to remember that the odds corresponds to the ratio between a probability and its complement to 1; so, for a priori probability equal to 0.95 very unfavorable for the null hypothesis of being true, the *a priori odds* is 0.95/0.0.05 = 19 or for a priori probability very favorable for the null hypothesis of being true equal to 0.90 the a priori odds is 0.90/0.0.10 = 9. Furthermore, we obtain an odds value of 1 for a probability of 0.5 and of 0.33 for a probability of 0.25, respectively.

Then, by multiplying the *prior odds* by the Bayes Factor, it is possible to calculate the posterior odds that, for an easy reading, can be converted in a probability value by remembering that p = odds / (1 + odds). For example, with BF of 0.2, 0.1, 0.05, and, finally, of 0.01 the above *prior odds* of 19 against the null hypothesis give *posterior odds* of 3.8, 1.9, 0.95, and 0.19. It is straightforward to obtain the corresponding posterior probability values of 0.792, 0.655, 0.487, and 0.159. Again, for the above *prior odds* of 9, we obtain *posterior odds* of 1.8, 0.9, 0.45, and 0.09 with the corresponding posterior probability values of 0.643, 0.474, 0.310, and 0.083. It has to be said that the above Bayes factor values correspond, according to Goodman [34] to a "Strength of Evidence" "weak", "moderate", "moderate to strong", and, finally "strong to very strong", respectively. Apart from considering some particular values of the Bayes Factor as shown before, it is very useful to consider that in the case of statistical tests based on the Gaussian distribution, as usually happens in the biomedical research, the "minimum Bayes Factor" is obtained by:

$$exp\left(-z^2/2\right) \qquad (3)$$

Where "z" is the quantile of the standardized Gaussian distribution corresponding to the obtained p-value; for instance z = 1.28155 for a p-value = 0.90, z = 1.6448 for a p-value of 0.95, z = 1.88079 for a p-value = 0.97, z = 1.95996 for a p-value = 0.975, z = 2.32635 for a p-value of 0.99, and, finally, z = 3.09023 for a p-value of 0.999. It has to be noted that the "Minimum Bayes Factor" corresponds to the strongest Bayes factor against the null hypothesis. Unfortunately, in Table 1 of the Goodman's paper [34] it has not reported that the probability values shown on the first column under the heading "P Value (Z Score)" has to be considered as *two tailed*. So, the values of the Minimum Bayes Factor shown on the second column are, obviously, only correct for the two tailed probability value obtained by dividing by two the values shown in the first column. In any case, a substantial decrease of the probability of the null hypothesis of being true has been obtained for all the situations shown in the table.

Table 1: Values of the posterior probabilities of the null hypothesis (H_0) of being true according to some values of the a priori odds calculated from the p-value obtained from the experimental data. Values of the "Minimum Bayes Factor (M-BF)" and of the "Symmetrical Bayes Factor" (S-BF) are shown.

P-value	Zeta	Minimum Bayes Factor	A priori p(H_0)	A priori Odds(H_0)	M-BF Posterior Odds (H_0)	M-BF Posterior p(H_0)	Symmetrical Bayes Factor	S-BF Posterior Odds (H_0)	S-BF Posterior p(H_0)
0.1	1.2816	0.4399	0.75	3	1.31973	0.56891	0.62591	1.87772	0.6525
0.1	1.2816	0.4399	0.5	1	0.43991	0.30551	0.62591	0.62591	0.38496
0.1	1.2816	0.4399	0.17	0.2048	0.0901	0.08265	0.62591	0.1282	0.11363
0.05	1.6449	0.2585	0.75	3	0.77557	0.4368	0.40716	1.22149	0.54985
0.05	1.6449	0.2585	0.5	1	0.25852	0.20542	0.40716	0.40716	0.28935
0.05	1.6449	0.2585	0.26	0.3514	0.09083	0.08327	0.40716	0.14306	0.12515
0.03	1.8808	0.1706	0.75	3	0.51167	0.33848	0.28595	0.85786	0.46175
0.03	1.8808	0.1706	0.5	1	0.17056	0.14571	0.28595	0.28595	0.22237
0.03	1.8808	0.1706	0.33	0.4925	0.08401	0.0775	0.28595	0.14084	0.12346
0.01	2.3264	0.0668	0.75	3	0.20042	0.16696	0.12518	0.37554	0.27302
0.01	2.3264	0.0668	0.5	1	0.06681	0.06262	0.12518	0.12518	0.11125
0.01	2.3264	0.0668	0.6	1.5	0.10021	0.09108	0.12518	0.18777	0.15809
0.001	3.0902	0.0084	0.75	3	0.02532	0.02469	0.01878	0.05633	0.05333
0.001	3.0902	0.0084	0.5	1	0.00844	0.00837	0.01878	0.01878	0.01843
0.001	3.0902	0.0084	0.92	11.5	0.09706	0.08847	0.01878	0.21594	0.17759

Table 1 of the Goodman's paper [34] has to be corrected as the following Table 1 shows for the part on the left regarding the Minimum Bayes Factor column and those that the M-BF Posterior odds and M-FB Posterior p(H_0). It has to be stressed that, in this case, the five values of the minimum Bayes Factor shown in Table 1 have been defined as a "weak", "moderate", "moderate", "moderate to strong", and "strong to very strong". It is an obvious consideration that a Bayes Factor equal to 1/10 for the null hypothesis against to the alternative hypothesis, it means that these study results have decreased the relative odds of the null hypothesis by 10-fold. Furthermore, it has also to consider that, the minimum Bayes factor described above does not involve a prior probability distribution over the non-null hypotheses and, consequently, it is a global minimum for all prior distributions. However, there is also a simple formula for the minimum Bayes factor in the situation where the prior probability distribution is symmetric and descending around the null value. This is given by:

$$exp\left(1\right) \cdot p \cdot ln\left(p\right) \qquad (4)$$

Where p is the p-value associated to the statistical significance obtained from the experimental data[3]. This symmetrical Bayes factor has been used in Nuzzo's paper [30] the last three columns of the above Table 1 report the values of the "minimum Bayes Factor for a symmetric prior probability distribution (Symmetrical); it has to be noted that the decrease of the probability is lower than that obtained by the minimum Bayes factor. Finally, it has also to mention the "objective" posterior probabilities that can be obtained, according to Jeffreys as reported by Berger [51]. For a Bayes factor calculated according to the above equation 2, the probabilities for H_0 and H_1 are respectively:

$$P(H_0 \mid x_{OBS}) = \frac{BF(x_{OBS})}{1 + BF(x_{OBS})}$$
$$P(H_1 \mid x_{OBS}) = \frac{1}{1 + BF(x_{OBS})} \qquad (5)$$

The above expressions are pertinent to the case of a prior probability equal to 0.5, leading to calculated prior odds of 1. So, the posterior odds are just equal to the Bayes factor and then, it has to apply the usual formula already shown for obtaining a probability value from odds. The formula for calculating the posterior probability of the alternative hypothesis (p (H_1)) is obtained by considering that it has to be use the reciprocal of the Bayes factor calculated for the null hypothesis. By substituting $BF_{(xOBS)}$ with 1/$BF_{(xOBS)}$ in the first formula the second is easily obtained. To this regards, it can be said that a p-value of 0.005 gives M-BF = 0.03625 and a S-MF = 0.07201 instead of 0.25852 and 0.40716 respectively obtained by a p-value = 0.05. This fact has an important consequence on the posterior H_0 probability of being true that for a p-value = 0.005 remains <0.05 until a prior p(H_0) ≤0.55 (at least) against H_0 instead of p(H_0) ≤0.15 (at least) in the case of p = 0.05. For sake of clarity, the above calculations have been done with the values of the minimum Bayes factor and with an increase of p(H_0) in steps of 0.05 from 0.05 to 0.95. Finally, it has to be noted that a probability against H_0 is actually in favor until a value of 0.5

The case of a greater significance level

Phase II clinical trials in Oncology tend to consider higher significance levels (ranging from 0.05 to 0.20) for reducing the number of the patients to be enrolled and, consequently for having a faster screening of the drugs potentially interesting for being tested for efficacy in a larger Phase III trial [54]. Furthermore, according to Jung [55], phase II trials in order to lower the sample size "use a surrogate outcome rather than a confirmatory endpoint and one-sided α of 0.05 to 0.20 and a power of 0.80 to 0.90, compared to two-sided α of 0.05 and a power of 0.90 or higher in phase III trials". Also the threshold for declaring a pharmacodynamics effect, as in the Phase 0 trials, is preferably put at 0.10 [56]; it has to be considered that Phase 0 trials allow to establish feasibility and to refine the trial methodology for anticancer drugs in a limited number of patients before a large number of patients are exposed to toxic doses of the study agent.

The importance of reducing the number of the patients to be enrolled in Phase II trials is well documented also by the proposal of Khan, Sarker and Hackshaw [57] consisting in accepting an "α level that is 'around' 10% and a power 'around' 80%", by exploiting the sawtooth behavior of the α and power function of the exact binomial statistical test [58]. For example, for demonstrating a difference from 0.10 to 0.20 with significance level (α) of 0.05 and power (1-β) of 0.80, the calculated sample size is of 78 subjects with 13 successes as the critical number for not accepting H_0. However, owing to the above mentioned sawtooth behavior, the actual values of α and 1-β are 0.0453 and 0.8081, respectively. Moreover, by accepting α = 5.67% and power = 77.7%, both close to the required levels of 0.05 and 0.80, the sample size is of 65 with a relevant saving of 13 subjects (16.7%).

A useful and exhaustive review on Phase II designs is from Mariani & Marubini [59]; this review is relevant also from the historical point of view since it summarizes all the main statistical methodology until the year of its publication. Finally, it has to be remembered, almost like a curiosity, that the FDA guidance [60] reports that the Center for Veterinary Medicine "generally considers a significance level of α = 0.10 useful as a conservative screen for identifying potential treatment-related safety concerns among endpoints in Target Animal Safety studies". In addition, also "Pairwise mean comparisons between each treatment against the control group are also performed using an unadjusted α = 0.10." So, as a conclusion, in preliminary trials of anticancer drugs the proposal of lowering the significance threshold seems rather questionable and problematic.

The sample sizes aspect

It is obvious that moving the significance threshold from 0.05 to 0.005 there is an important increase in the sample sizes necessary to be enrolled in a trial, keeping fixed the other ingredients of the sample size calculation that are the effect size or difference and variability for continuous variables, the difference and the baseline proportion for qualitative variables, the power and the statistical significance test. The paper from Benjamin et al. [1] reports that "for a wide range of common statistical tests, transitioning from a P-value threshold of α = 0.05 to α = 0.005 while maintaining 80% power would require an increase in sample sizes of about 70%".

It is worthwhile to underline that it could be argued that the switching from 0.05 to 0.005 actually refer to a switch from 0.025 to 0.0025 since the ICH E guideline [61] refers to a two-sided statistical test. It had to be noted that in the case of a sample size calculation for an unpaired Student's t test, power of 0.80 and effect size values ranging from 0.25 to 2.5 by step of 0.01, the increase of the sample size is globally of about the 65.97% for a power of 0.80. Then there is a decrease for increasing values of the power; for instance, it is of 63.62% for a power of 0.85 and becomes of 59.60% for a power of 0.90, as a further demonstration of the non-linearity relationships between the two functions of the statistical significance (α) and of the power (1-β). In any case the sample size increase has to be judged as relevant and, maybe, not ethically acceptable given the current limitations of the number of patients who can be actually enrolled in clinical trials and of the economic resources available.

Benjamin et al. [1] recognized that only fewer studies could be effectively conducted using current experimental designs and budgets. Furthermore in Figure 2, they showed the benefit of this p-values switching and its consequences; particularly, they stated without any further explanation that the "false positive rates would typically fall by factors greater than two". Then, Benjamin et al. [1] concluded with a series of documented claims such as "Increasing sample sizes is also desirable because studies with small sample sizes tend to yield inflated effect size estimates [62], and publication and other biases may be more likely in an environment of small studies [63]" and self-citations such as "considerable resources would be saved by not performing future studies based on false premises" and "We believe that efficiency gains would far outweigh losses" that, of course, have to be demonstrated. In any case, the huge increase in sample size calculation has a dramatically economic impact, and, above all, a series of ethical consequences that have to be appropriately considered and resolved.

An intriguing case

Recently Combes et al. [64,65] published on a top medical journal an international controlled trial comparing venovenous extracorporeal membrane oxygenation (ECMO) with the usual standard of care, but allowing for the patients in the control group the crossover to ECMO if they had refractory hypoxemia. The primary end point was mortality at 60 days. The key secondary end point was treatment failure, which was defined as crossover to ECMO or death in patients in the control group and as death in patients in the ECMO group. It is very well known that the acute respiratory distress syndrome (ARDS) is a very severe disease associated with a high mortality exceeding 60%. Then it is very understandable the expectation that this trial had aroused in the medical world, particularly in physicians working in the Intensive Care Unit.

The sample size calculation was based on a very sophisticated statistical methodology such as group sequential analysis, triangular test, etc. that it is not possible to comment in depth here. However it has to be said that the trial had the ambitious aim of demonstrating a 20% reduction in the expected mortality at 60 days (60% in the group receiving conventional ventilation vs. 40% among those receiving early ECMO support). Accordingly, it has be stated: "for a 80% power, at an alpha level of 5% and with a group sequential analysis occurring after the randomization of every 60 participants, the maximum sample would need to be 331 participants." Furthermore the statistical analysis was very complicated by the fact that 28% of the patients in the control group crossed over to ECMO for refractory hypoxemia. About this it has to report what the authors very correctly wrote "We were aware of this potential problem when we started the trial, but many investigators felt that it would have been unethical to prohibit crossover to ECMO in patients with very severe hypoxemia".

Unlikely, the statistical analysis on the primary end point at 60 days showed a relative risk of 0.76; 95% confidence interval [CI], 0.55 to 1.04; P = 0.09. Also an additional statistical analysis (log-rank test), actually carried out according to a not justifiable criterion in my opinion, showed a result not statistically significant: "the hazard ratio for death within 60 days after randomization in the ECMO group, as compared with the control group, was 0.70 (95% CI, 0.47 to 1.04; P = 0.07)" Finally, also a multivariable analysis gave not statistically significant results, as the authors wrote: "Adjustment for important prognostic factors did not change the results." However, the fact that it is has not clearly stated how these results have been obtained and, consequently, that they cannot be reproduced is, in my opinion, particularly disturbing. For example if we carry out a simple χ^2 analysis of the 44/124 vs. 57/125 proportions of events in the ECMO and control group, respectively as the Table 1 of Combes et al. [64] shows, we obtain: Chi-Square = 2.6423 with p = 0.1041 and a Continuity Adjusted Chi-Square = 2.2393 with p = 0.1345, very different from the p = 0.09 reported. Finally, at the Fisher's exact test the two-tailed p is 0.1217. Of course, also the relative risk is different: 0.7782 95%CI: 0.5736 - 1.0556 instead of: 0.76 (0.55 to 1.04) shown in Table 1.

Furthermore, even if the secondary end points turned out to be statistically significant in favour of the ECMO treatment ("the relative risk of treatment failure, defined as death by day 60 in patients in the ECMO group and as crossover to ECMO or death in patients in the control group, was 0.62 with 95% Confidence Interval of: 0.47 to 0.82; P<0.001, for example), the authors had to sadly and sharply write that: "In conclusion, the analysis of the primary end point … showed no significant benefit of early ECMO, as compared with a strategy of conventional mechanical ventilation, which included crossover to ECMO (used by 28% of the patients in the control group)." The impact of this result that, according to the Evidence Based Medicine (EBM), does not allow to recommend the ECMO treatment in these very severely ill patients is, of course, very frustrating for the physicians working in the Intensive Care Units. So, the question that arises almost spontaneously is whether a difference of a few cents (4 or 2, depending on the statistical test carried out and on the exact at the fourth decimal figure p-values obtained) should be considered so relevant as to make inconclusive such a clinically important result.

To this regards, it has to do some clarifications. Firstly, it is often misunderstood that the statistical significance threshold of 0.05 has to be always considered, in the clinical trials settings as two-tailed, and, consequently the significance threshold is of 0.025; so the difference is of 65 or 45 thousandths since the statistical significance in the paper [64] has been reported only at the second decimal figure. In any case, if the statistical significance threshold had been settled at 0.10 during the planning of the study, would not have had the current problems in the interpretation of its results and in accepting an innovative strategy of treatment. Secondly, it has to critically reconsider the rigid position of the regulatory authorities to judge a controlled clinical trial as inconclusive if the primary outcome has not been demonstrated by means of a statistically significant result. Even if this position can be considered as justifiable for trials aimed to a drug registration for its commercialization, I think that it has to be assumed a more flexible attitude in the case of a treatment such as the ECMO in the Intensive Care settings. Indeed, it has also to consider:

a) The clinical context in which the trial has been carried out;

b) The potential for care of the current treatment;

c) The methodological statistical aspects such as the real difficulties in doing a direct and easy comparison owing to the crossover from the control to the experimental group (or, generally, a crossover even for both the treatment groups);

d) The limitations of the trial that have been clearly and exhaustively reported by Combes et al. [64] at the end of the paper; and, lastly,

e) Some pitfalls in the planning of the controlled clinical trials those subsequent amendments (this trial had as many as ten amendments) try to fix more or less successfully and the remarkable duration of the trial that was approved in 2010 and published 8 years after. So, I think that it is possible to consider the trial as adequately supportive of the ECMO treatment [65].

Conclusion

The recent proposal of moving down to 0.005 the statistical significance threshold is, of course, well-motivated in the Benjamin et al. [1] and also in the previous paper from Johnson [13]. However, it has to say that accepting such a proposal is involves such a change in the scientific world, in the mentality of researchers, in drug development by the pharmaceutical companies that could have negative consequences at least in the first years following. I think that it is mandatory that researchers have an adequate knowledge of the statistical method and also of the meaning of the p-values in order to appropriately consider the results of the research and to be absolutely aware of their use. One could begin to request that p-values be accompanied by considerations about the probability that the null hypothesis (and / or the alternative) is true. These considerations should have an appropriate prominence perhaps even in the context of the conclusions of the abstract of the published papers.

References

1. Benjamin DJ, Berger JO, Johnson VE (2018) Redefine statistical significance. Nature Human Behaviour 2: 6-10.
2. Everitt BS, Skrondal A (2010) The Cambridge Dictionary of Statistics. 4th (Edn.) Cambridge University Press, New York, USA.
3. Ioannidis JPA (2018) The Proposal to Lower P Value Thresholds to .005. JAMA 319(14): 1429-1430.
4. Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Medicine 2(8): e124.
5. Goodman S, Greenland S (2007) Why most published research findings are false: Problems in the analysis. PLoS Medicine. 4(4): e168.
6. Goodman S, Greenland S (2018) Assessing the unreliability of the medical literature: A response to "Why most published research findings are false". 2007 Johns Hopkins University, Department of Biostatistics.
7. Berger JO, Sellke T (1987) Testing a Point Null Hypothesis: The Irreconcilability of P-Values and Evidence. Journal of the American Statistical Association 82: 112-122.
8. Goodman SN (1994) Confidence limits vs. power calculations. Epidemiology 5(2): 266-268.
9. Moran JL, Solomon PJ (2004) A farewell to P-values. Critical Care and Resuscitation 6(2): 130-137.
10. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, et al. (1989) The Empire of Chance: How Probability Changed Science and Everyday Life. New York, USA.
11. Morrison DE, Henkel RE (1969) The significance test controversy-a reader. Chicago, USA.
12. Harlow LL, Mulaik SA, Steiger JH (1997) What if there were no significance tests. Hillsdale NJ: Lawrence Erlbaum Associates.
13. Johnson VE (2013) Revised standards for statistical evidence. Proceedings of the National Academy of Sciences 110(48): 19313-19317.
14. Gelman A, Robert CP (2014) Revised evidence for statistical standards. Proceedings of the National Academy of Sciences 111(19): E1933.
15. Gaudart J, Huiart L, Milligan PJ, Thiebaut R, Giorgi R (2014) Reproducibility issues in science, is P value really the only answer? Proceedings of the National Academy of Sciences 111(19): E1934.
16. Johnson VE (2014) Reply to Gelman, Gaudart, Pericchi: More reasons to revise standards for statistical evidence. Proceedings of the National Academy of Sciences 111(19): E1936-E1937.
17. Jager LR, Leek JT (2014) An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 15(1): 1-12.
18. Benjamini Y, Hechtlinger Y (2014) Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. Biostatistics 15: 13-16.
19. Cox DR (2014) Discussion: Comment on a paper by Jager and Leek Biostatistics 15: 16-18.
20. Gelman A, Orourke K (2014) Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. Biostatistics 15(1): 18-23.
21. Goodman SV (2014) Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 15: 23-27. https://doi.org/10.1093/biostatistics/kxt035.
22. Ioannidis JP (2014) Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. Biostatistics 15(1): 28-36. https://doi.org/10.1093/biostatistics/kxt036.
23. Schuemie MJ, Ryan PB, Suchard MA, Shahn Z, Madigan D (2014) Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 15(1): 36-39.
24. Wasserstein RL, Lazar NA (2016) The ASA's statement on P-values: context, process, and purpose. The American Statistician 70(2): 129-133.
25. Gattinoni L, Giomarelli P (2015) Acquiring knowledge in intensive care: merits and pitfalls of randomized controlled trials. Intensive Care Medicine 41(8): 1460-1464.
26. Cesana BM (2016) Negative randomized clinical trials (RCTs): further insight from the biostatistician's point of view. Intensive Care Medicine 42(1):136.
27. Trafimow D (2014) Editorial. Basic and Applied Social Psychology 37(1):1-2.
28. Trafimow D, Michael M (2015) Editorial Basic and Applied Social Psychology 37(1):1-2.
29. Nature | Research Highlights: Social Selection. Journal bans P values. Nature 5 March 2015;519:9 doi:10.1038/519009f.
30. Nuzzo R (2014) Statistical Errors P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506(7487): 150-152.
31. Editorial (2014) Number crunch the correct use of statistics is not just good for science - it is essential. Nature 13 February 506(7487): 131-132.
32. Goodman SN (1992) Replication, p-values and evidence. Statistics in Medicine 11(7): 875-879
33. a. Goodman SN (1999) Toward evidence-based medical statistics 1: The P value fallacy. Annals of Internal Medicine 130(12): 995-1004. 33b.
34. Goodman SN (1999) Toward evidence-based medical statistics 2: The Bayes Factor. Annals of Internal Medicine 130(12): 1005-1013.
35. Goodman SN (2001) Of P-Values and Bayes: A Modest Proposal. Epidemiology 12(3): 295-297.
36. Yaddanapudi LN (2016) The American Statistical Association statement on P-values explained. Journal of Anaesthesiology Clinical Pharmacology 32(4): 421-423.
37. Nuzzo R, Johnson V, Senn S (2018) ASA statement on P-values and statistical significance: Development and impact.
38. Mark DB, Lee KL, Harrell FR (2016) Understanding the Role of P Values and Hypothesis Tests in Clinical Research. Special Communication JAMA Cardiology 1(9): 1048-1054.
39. Figueiredo Filho DB, Paranhos R, da Rocha EC, Batista M, da Silva JA, et al. (2013) When is statistical significance not significant?. Brazilian Political Science Review 7(1): 31-55.
40. https://www.youtube.com/watch?v=eyknGvncKLw
41. https://www.youtube.com/watch?v=128yz0OCG-I
42. https://www.youtube.com/watch?v=-MKT3yLDkqk
43. http://www.youtube.com/watch?v=ax0tDcFkPic&feature=related
44. Fisher RA Statistical Methods for Research Workers. 1925. Oliver & Boyd. Edinburgh: ISBN 0-05-002170-2.
45. Fisher RA (1935) The Design of Experiments (9th Edn. 1971) (1st Edn. 1935) Macmillan. ISBN 0-02-844690-9.
46. Fisher RA Sir, Yates F (1938) Statistical tables for biological, agricultural and medical research. (6th Edn.). 1953 (1st Edn.). Oliver and Boyd London, UK.
47. Hubbard R, Bayarri MJ (2003) Confusion over measures of evidence (p's) versus errors (a's) in classical statistical testing. The American Statistician 57(3): 171-182.

48. Goodman SN (1993) P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. American Journal of Epidemiology 137(5): 485-496.

49. Gibbons JD, Pratt JW (1975) P-values: Interpretation and Methodology. The American Statistician 29(1): 20-25.

50. De Groot MH (1973) Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. Journal of the American Statistical Association 68: 966-969.

51. Christensen R (2005) Testing Fisher, Neyman, Pearson, and Bayes. The American Statistician 59(2): 121-126.

52. Berger JO (2003) Could Fisher, Jeffreys and Neyman Have Agreed on Testing? Statistical Science 18(1): 28-32.

53. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, et al. (2016) Statistical tests, Pvalues, confidence intervals, and power: a guide to misinterpretations European Journal of Epidemiology 31(4): 337-350.

54. Chow SC, Wang H, Shao J (2008) Sample Size Calculations in Clinical Research. (2nd Edn.) Chapman and Hall/CRC Taylor & Francis Group Boca Raton.

55. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, et al. (2001) Clinical Trial Designs for Cytostatic Agents: Are New Approaches Needed? Journal of Clinical Oncology 19(1): 265-272.

56. Jung SH (2013) Randomized Phase II Cancer Clinical Trials. 2013 Chapman and Hall/CRC Taylor & Francis Group Boca Raton.

57. Crowley J, Hoering A (2012) Handbook of Statistics in Clinical Oncology (3rd Edn.) 2012 CRC Press Taylor & Francis Group Boca Raton.

58. Khan I, Sarker SJ, Hackshaw A (2012) Smaller sample sizes for phase II trials based on exact tests with actual error rates by trading-off their nominal levels of significance and power. British Journal of Cancer 107: 1801-1809.

59. Cesana BM, Reina G, Marubini E (2001) Sample Size for Testing a Proportion in Clinical Trials: A "Two-Step" Procedure Combining Power and Confidence Interval Expected Width. The American Statistician 55: 288-292.

60. Mariani L, Marubini E (1996) Design and Analysis of Phase II Cancer Trials: A Review of Statistical Methods and Guidelines for Medical Researchers. International Statistical Review 64(1): 61-88.

61. FDA Guidance for Industry Target Animal Safety Data Presentation and Statistical Analysis p. 226.

62. ICH Harmonised Tripartite Guideline Statistical Principles for Clinical Trials E9 Current Step 4 version.

63. Gelman A, Carlin J (2014) Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. Perspect Psychol Sci 9(6): 641-651.

64. Combes A, Hajage D, Capellier G, Demoule A, Lavoué S, et al. (2018) Extracorporeal Membrane Oxygenation for Severe Acute Respiratory Distress Syndrome. New England Journal of Medicine 378(21): 1965-1975.

65. Fanelli D, Costas R, Ioannidis JPA (2017) Meta-assessment of bias in science. Proc Natl Acad Sci USA 114(14): 3714-3719.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2018.06.001359

Bruno Mario Cesana. Biomed J Sci & Tech Res