La Medicina del Lavoro

Med Lav 2017; 108, 5: 327-331
DOI: 10.23749/mdl.v108i5.6603

Health significance and statistical uncertainty. The value of P-value

Dario Consonni[1], Pier Alberto Bertazzi[2]

[1] Epidemiology Unit, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
[2] Post-Graduate School of Occupational Medicine, University of Milan, Milan, Italy

Key words: Statistics; p-value; confidence interval; environmental and occupational health

Parole chiave: Statistica; valore p; intervallo di confidenza; medicina del lavoro e dell'ambiente

Summary

Background: *The P-value is widely used as a summary statistics of scientific results. Unfortunately, there is a widespread tendency to dichotomize its value in "P<0.05" (defined as "statistically significant") and "P>0.05" ("statistically not significant"), with the former implying a "positive" result and the latter a "negative" one.* **Objective**: *To show the unsuitability of such an approach when evaluating the effects of environmental and occupational risk factors.* **Methods:** *We provide examples of distorted use of P-value and of the negative consequences for science and public health of such a black-and-white vision.* **Results**: *The rigid interpretation of P-value as a dichotomy favors the confusion between health relevance and statistical significance, discourages thoughtful thinking, and distorts attention from what really matters, the health significance.* **Discussion**: *A much better way to express and communicate scientific results involves reporting effect estimates (e.g., risks, risks ratios or risk differences) and their confidence intervals (CI), which summarize and convey both health significance and statistical uncertainty. Unfortunately, many researchers do not usually consider the whole interval of CI but only examine if it includes the null-value, therefore degrading this procedure to the same P-value dichotomy (statistical significance or not).* **Conclusions**: *In reporting statistical results of scientific research present effects estimates with their confidence intervals and do not qualify the P-value as "significant" or "not significant".*

Riassunto

«Il valore di P-value. Incertezza statistica e rilevanza sanitaria». **Background**: *Il valore P è ampiamente utilizzato come indice statistico riassuntivo dei risultati scientifici. Sfortunatamente, c'è una diffusa tendenza a dicotomizzare il suo valore in "P<0.05" (definito come "statisticamente significativo") e "P>0.05" ("statisticamente non significativo"), con l'implicazione che nel primo caso il risultato sia "positivo" (cioè che la associazione – negativa o positiva che sia – esista) e "negativo" nel secondo.* **Obiettivo**: *Mostrare i limiti e la inappropriatezza di un tale approccio per la valutazione dei fattori di rischio occupazionali e ambientali.* **Metodi**: *Vengono presentati esempi sull'uso distorto del valore P e delle conseguenze negative di questo visione "bianco o nero".* **Risultati**: *La rigida interpretazione del valore P come una dicotomia favorisce la confusione tra rilevanza sanitaria e significatività statistica, scoraggia il pensiero critico e distoglie l'attenzione da ciò che realmente conta, la rilevanza sanitaria.* **Discussione**: *Un modo molto migliore di esprimere e comunicare i risultati scientifici consiste nel riportare le stime dell'effetto (ad esempio, rischi, rapporti fra rischi o differenze tra rischi) e i loro intervalli di confidenza, che insieme sintetizzano e forniscono sia la rilevanza per la salute sia l'incertezza statistica. Sfortunatamente, molti ricercatori non considerano l'intero*

Pervenuto il 11.7.2017 - Accettato il 12.10.2017

Corrispondenza: Dario Consonni, UO Epidemiologia, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Via San Barnaba, 8 - 20122 Milano

open access www.lamedicinadellavoro.it

intervallo, ma esaminano solo se esso contiene oppure no il valore nullo, in questo modo degradando questa procedura alla stessa dicotomia del valore P (significatività statistica o no). **Conclusioni**: *Quando si riportano i risultati statistici di ricerche scientifiche, presentare le stime di effetto con i loro limiti di confidenza e non qualificare il valore P come "significativo" o "non-significativo".*

Background

Probability has a fundamental role in various scientific disciplines. For example, probability of disease is used when calculating risks, rates, hazards, odds, and their ratios ("relative risks", RR) or differences. However, there is a single use of probability that stands out and pervades observational, etiologic, and evaluation studies: the "P-value", often also referred to simply as "P" or "p". This statistics is widely used to summarize results, to make inference, and finally draw conclusions. However, due to inherent technical limits and, most importantly, to the misinterpretation by many users, P-value turns out to be an incomplete, unsatisfactory, and misleading means of summarizing results.

The literature discussing limitations and misuse of P-value is vast (2, 4-8, 10-15). In this paper, we do not discuss the well-known statistical limits of P-value. Instead, we provide some examples on the distorted use of P-value and of the negative consequences for science and public health of an approach based on this single statistics. Then we discuss a simple and better alternative, the confidence interval, conditional to its correct use. Finally, we conclude by reiterating few simple recommendations in reporting statistical results. We focus our attention on public health domains, including occupational, environmental, and clinical fields. Other scientific fields (e.g., -omics) may require a somewhat different and more elaborate discussion.

The P-value

What is a P-value? Its correct definition is not simple for non-statisticians. Informally, P-value is "the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." (15). Very often the statistical model of choice is the so-called "null hypothesis" (e.g., that there are no differences between two groups), so that most statistical inference is based on the so-called "null hypothesis significance testing" (NHST) even though other hypotheses might be more relevant (5, 7).

Many researchers would probably have difficulties in grasping in depth the above concept and probably ignore the numerous theoretical assumptions and subtleties implicit in P-value calculation (5). However, every day the same researchers make use of P-values in analyzing their data, in drawing conclusions from their analyses, and in interpreting scientific papers. Therefore, P-values are largely subject to misinterpretation. Notwithstanding repeated warnings over the last decades about the theoretical and technical limits of P-value and, most importantly, against its abuse and misuse, too often the conclusions of a scientific study are based on this single summary of statistical analyses.

Quite a few books (8, 10, 12) and papers (2, 4-7, 11, 13-15) discuss the problems in using and interpreting P-values. We deem that the *single* most pernicious misuse is the widespread tendency of dichotomizing the P-value in "P<0.05" (called "statistically significant") and "P>0.05" (qualified as "statistically not significant", sometimes abbreviated as "NS"). (P=0.05 is included in either one or the other category). Frequently, studies with P<0.05 (being the exact value, say, P=0.0001 or P=0.04) are incorrectly labelled as "positive", implying that the investigated association does exist. By the same token, other studies yielding P>0.05 (again, being the exact value, say, P=0.06 or P=0.90) are incorrectly regarded as "negative", implying that the apparent association is not real (8). This behavior was named "dichotomania" (5). A few examples of the negative consequences of such a black-and-white vision for science and public health follow.

A study among workers exposed to a suspect carcinogen yielding RR=1.2 for lung cancer mortality with P=0.04 would be regarded as "positive". A similar study with RR=2.0 and P=0.06 would be labelled as “negative”, although the effect is much larger. The dichotomized P-value favors in this way the confusion between “health significance” and “statistical significance”. (Sometimes, a P=0.06 would be rescued by calling it “borderline”, but remains statistically not significant in the mind of many). This is particularly true when we analyze big datasets: since the P-value largely depends on sample size, most P-values will be very low (say, P<0.0001), even when the effects are quite small not to say trivial (10).

Too often researchers rely on the P-value from a single study to conclude about the existence of an association. However, knowledge in science advances through replication of results under different conditions. In most situations, a single study cannot lead to a conclusive evaluation (e.g., whether an agent is toxic) by itself. Imagine ten studies on a suspect carcinogen, some with RR>1.0, some with RR<1.0, some with P<0.05, and some with P>0.05. A systematic review, possibly with a meta-analysis pooling these results together, could calculate, say, a summary RR=1.4, with an overall P=0.001. Admittedly with the absence of important biases, this result conveys strong evidence in favor of the carcinogenicity of that agent, no matter how the authors of the individual studies (or their readers) labelled their results (statistically significant or not) (14).

In addition, the dichotomized P-value favors data dredging to obtain a “statistically significant” result (so-called “P-value chasing” or “P-hacking”) as, for example, with the use of several statistical tests (e.g., parametric *vs* non-parametric), multiple testing (e.g., subgroup analyses), or fitting of several multivariable models with different sets of confounders. Indeed, as already noted, P-values between 0.041 and 0.049 occurs too often in scientific papers (3). This malpractice is not solely responsibility of the researchers, because their work has to cope with a “culture that selectively publishes or otherwise focuses on statistically significant results” (7): in fact, many peer reviewers, journal editors, and journal readers continue to regard P<0.05 as a “positive” result and P>0.05 as a “negative” one. These compulsive behaviors in search of statistical significance discourage thoughtful thinking and distort attention from what really matters, i.e., the health significance.

Moreover, selection of “statistically significant” results induces false expectations such as a high reproducibility of results (when instead reproducibility depends on study power), and overestimation of effects (7).

Diagnostic process and scientific research

For health professionals we would like to point out some similarities between clinical and statistical tests, and between diagnostic process and scientific research in general. Consider a clinical laboratory test, for example two blood glucose measurements in two patients, 109 and 111 mg/dL, with threshold of normality set at, say, 110 mg/dL. Although the second result would have an asterisk marked on the lab sheet, we think every physician would consider the two as equivalent (they provide similar – only suggestive – strength of evidence for diabetes) and ask for further diagnostic tests for both patients. On the contrary, two patients with 200 and one with 111 mg/dL (both over the threshold and with asterisks) would certainly be considered quite differently (the former provides stronger evidence for diabetes).

The same (should) hold for a statistical test. P-values should be evaluated in a less rigid and more “qualitative” way: P=0.04 and P=0.06 (one below and one over the conventional threshold, the first perhaps with an asterisk in the statistical software output) should be considered as equivalent (they provide similar – only suggestive – strength of evidence against the null hypothesis). Conversely, P=0.0001 and P=0.04 (both below the threshold and with asterisks) should be regarded as different (the former provides stronger evidence against the null) (5, 14).

More in general, in caring for ill persons, physicians ask for and look at many diagnostic tests, but their conclusions do not stem from the result of a single test. Instead, to reach a diagnosis they interpret and weigh the whole body of information at hand (clinical history and test results). Similarly, when studying a potentially toxic agent, scientists can and should perform several statistical analyses, but they should *not* base their conclusions on a single P-value. Rather, they should carefully consider information from several fields including biology, toxicology, epidemiology, and the like.

Confidence intervals

A much better way to express and communicate scientific results involves reporting *effect estimates* (e.g., risks, risks ratios or risk differences) and their *confidence intervals* (CI). Without focusing on the null hypothesis, this "triad" of numbers (the effect estimate and the lower and upper confidence limit) conveys both *health significance* and *statistical uncertainty*: the larger the sample size, the narrower the CI, the higher the precision (i.e., the less the uncertainty) (11, 12). Indeed, meta-analyses make use of effect estimates and their CIs, not of statistical significance (12). Formulas based on CI are also very useful (and in our view more understandable to non-specialists than P-value-based formulas) in research planning phases for power and sample size calculations (1).

Usually a 95% confidence interval is calculated. There is a close correspondence between a 95% CI and P-value: if 95% CI includes the "null value" (e.g., a relative risk of 1.00 or a risk/mean difference of 0.00), then P>0.05; if 95% CI does not include the null value, then P<0.05. This has a very unfortunate consequence: most often, researchers do not look at the whole interval, but degrade the results as "statistically significant" or not, falling back in the P-value dichotomy (positive *vs* negative) (5, 7). Moreover, one confidence limit only (lower or upper) is looked upon. As an example, the result of a study with RR=1.50 and 95% CI=1.01-2.23 for lung cancer mortality among workers exposed to a carcinogen would be claimed as "statistically significant" (Figure 1, Study 1). Conversely, the same RR of 1.50 with 95% CI=0.99-2.27 in another study would be regarded as "statistically not significant", not considering that the upper limit (2.27) is compatible with a more than double cancer excess among the exposed (Figure 1, Study 2). For this reason, to discourage this simplistic black-and-white behavior, some suggest to report 90% CIs (9, 12, 14).

Note that in the run, and contrary to what many people think, a narrow CI is more important than a low P-value. For example, in a meta-analysis, an RR of 1.50 with 95% CI=0.90-2.50 (hence P>0.05, "statistically not significant") in a study (Figure 2, Study 1) will be given more weight (because of its larger sample size) than an RR of 5.00 with 95% CI=1.10-22.7 (hence P<0.05) in another study (Figure 2, Study 2) (11).

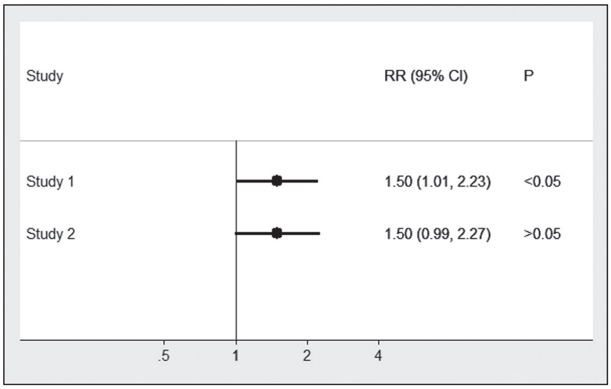

Figure 1 - Example of two studies with identical effect estimates (risk ratio, RR) and quite similar confidence intervals (CI), one with P<0.05 (Study 1) and one with P>0.05 (Study 2). The two studies are absolutely equivalent but they would be regarded as different (one "statistically significant" and the other not) based on the dichotomized P-value or on the fact that the lower confidence limit crosses the null value

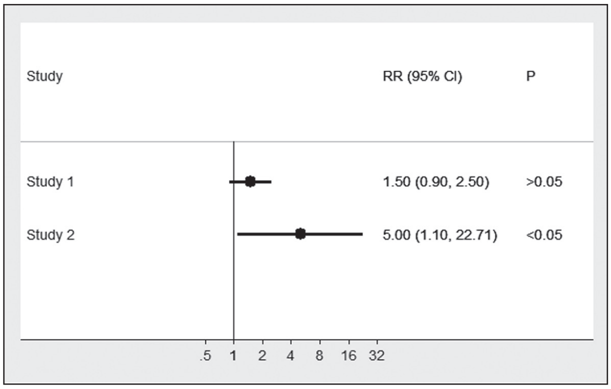

Figure 2 - Example of two studies, one with a smaller effect estimate (risk ratio, RR), a narrower confidence interval (CI), "statistically not significant" (Study 1, p>0.05 and CI that includes the null value) and the other with a larger RR, a wider CI, "statistically significant" (Study 2, P<0.05 and CI that exclude the null value). Study 1, being larger, would be given more weight in a meta-analysis

Conclusions

Dichotomized P-values are attractive, because they (appear to) simplify life. We understand the widespread tendency to compare P with the 0.05 threshold (we all have been taught to do so in statistical courses). However, it is time to dismiss this simplistic behavior (7). Science charges us with a more complex task than calculating a single number and comparing it to a completely arbitrary threshold.

As suggested, a strong cultural change (which includes abandoning the focus on hull hypothesis, NHST) is needed in the way statistics is taught, used, and interpreted (5, 7). While awaiting that time, we reiterate three simple suggestions to authors submitting manuscripts to scientific journals:

1. When possible, present effect estimates along with their confidence intervals (preferably at 90% level) instead of P-values; examine the whole interval (not just the lower or upper limit); and avoid to qualify the result as statistically significant or not based on the mere fact that the interval crosses (or not) the null value. An interval is an interval, not a point.
2. Do not write in the Methods section sentences like "We considered statistically significant a P<0.05".
3. If you report P-values, avoid labelling them as "statistically significant" or not; instead, evaluate them in a non-rigid, qualitative way and consider the health relevance of the findings (e.g., the extent of the estimated effect).

Several Journals already recommend doing so. In the "International Journal of Epidemiology" (IJE, the official journal of the International Epidemiological Association, IEA), the Instructions for Authors read: "*In the IJE we actively discourage the use of the term "statistically significant" or just "significant" and such statements in method sections as "findings at p<0.05 were considered significant". Where used, we ask authors to provide effect estimates with confidence intervals and exact P values, and to refrain from the use of the term "significant" in either the results or discussion section of their papers*" (https://academic.oup.com/ije/pages/Instructions_To_Authors). Similarly, "Epidemiology" in the author's instructions states: "*[…] we strongly discourage the use of categorized P-values and language referring to statistical significance […]. We prefer instead interval estimation, which conveys the precision of the estimate with respect to sampling variability.*" (http://edmgr.ovid.com/epid/accounts/ifauth.htm). Similar guidelines are reported in the "American Journal of Epidemiology" and "Occupational and Environmental Medicine". We hope other medicine and public health journals will join them in recommending better ways to report statistical results.

No potential conflict of interest relevant to this article was reported by the authors

References

1. Bland JM: The tyranny of power: is there a better way to calculate sample size? Br Med J 2009; 339: b3985
2 Consonni D, Bertazzi PA: La probabilità nelle scienze biomediche. Riv Ital Med Leg 2015; 4: 1449-1473
3 de Winter JCF, Dodou D: A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). PeerJ 2015; 3: e733
4 Goodman S: A dirty dozen: Twelve P-Value misconceptions. Semin Hematol 2008; 45: 135-140
5 Greenland S. Invited commentary: The need for cognitive science in methodology. Am J Epidemiol 2017; 186: 639-645
6 Greenland S, Senn SJ, Rothman KJ, et al: Statistical tests, P-values, confidence intervals, and power: A guide to misinterpretations. Eur J Epidemiol 2016; 31: 337-350
7 Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. Am J Epidemiol 2017; 186: 627-635
8 Hill AB: Principles of Medical Statistics (Eight Edition). London (UK): The Lancet Limited, 1967
9 Mensi C, De Matteis S, Catelan D, et al. Geographical patterns of mesothelioma incidence and asbestos exposure in Lombardy, Italy. Med Lav 2016: 107: 340-355
10 Miettinen OS: Theoretical Epidemiology. Principles of Occurrence Research in Medicine. New York (NY): Wiley, 1985
11 Poole C: Low P-values or narrow confidence intervals: Which are more durable? Epidemiology 2001; 12: 291-294
12 Rothman KJ, Greenland S, Lash TL: Modern Epidemiology, 3rd Edition. Philadelphia (PA): Lippincott Williams & Wilkins, 2008
13 Stang A, Poole C, Kuss O: The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010; 25: 225-230
14 Sterne JAC, Davey Smith G: Sifting the evidence - what's wrong with significance tests? Br Med J 2001; 322: 226-231
15 Wasserstein RL, Lazar NE: The ASA's statement on p-values: Context, process, and purpose. Am Stat 2016; 70: 129-133

Acknowledgements: *Several concepts presented here originated from discussions with Annibale Biggeri (University of Florence) at the Epidemiology Master Program, University of Turin. The way they were developed and expressed in this work are solely our responsibility.*