La Medicina del Lavoro

Med Lav 2018; 109, 2: 157-159

A short defence of p-value and statistical significance / *Una breve difesa del p-value e della significatività statistica*

Dear Editor,

the Commentary by Consonni and Bertazzi, recently published in "La Medicina del Lavoro" (1), raises the extremely important matter of the misuse of p-value and statistical inference in epidemiological studies. It represents a further contribution to a large campaign against the abuse of p-value and null hypothesis testing carried out by many other prestigious scientists (2, 3, 5, 8, 9). However, in several instances, some of the suggested remedies to this abuse can cause further insidious troubles. In particular, the use of confidence intervals (CI) instead of p-values can be treacherous when applied to categorical variables with more than two levels (polytomous variables).

Table 1 shows a simulated example in a hypothetical case-control study involving an exposure expressed by a categorical three-level variable. For reasons of simplicity confounding was not considered. Using a standard logistic regression model and assuming the A category as the referent, the following estimates of association were obtained (OR_1 in table 1): OR=0.72 for the B level, 95%CI: 0.54-0.95; OR=1.2 for the C level, 95%CI: 0.92-1.7. Results seem to indicate a small protective effect for subjects belonging to the B category and a very small excess risk (if any) for subjects exposed to the C level. As a whole, these results suggest that large differences in risk between the three categories are unlikely and, consequently, that no clear association was found between the considered disease and the studied risk factor. Let now hypothesize that another analysis was made on the same data by applying a similar statistical approach, but selecting the B category as the referent (OR_2 estimates in table 1). The following results were observed: OR=1.4 for the A category, 95% CI: 1.1-1.8; OR=1.7 for the C category, 95%CI: 1.2-2.4. Even if the two analyses include the same information, the pattern of risk seems now quite different and suggests that the considered risk factor is associated to the risk of developing the studied disease. The likelihood ratio test for the two models (that actually are two different parametrizations of the same model) provided the same result, namely: chi square=10.47 (with 2 degrees of freedom), p=0.005, pointing out a "highly statistically significant" association. It could be objected that the confusing pattern observed for the OR_1 estimates and their related 95%CI was caused by a wrong selection of the referent category. For instance, category B could have corresponded to the unexposed or very low exposed subjects and the C category to the "heaviest" exposed ones. However, in several actual situations this information is not available. For example, the three levels of exposure could correspond to three different jobs inside an industry producing some potentially toxic compound, in the absence of measures of environmental concentrations, a situation commonly encountered in historical cohort studies. The identification of an external (allegedly) unexposed group as referent often does not lead to a desirable solution, due to the very insidious combination of exposure misclassification bias and healthy worker effect (6). Another example involves the area of residence of subjects recruited in large population-based studies, which could be associated to a large set of different exposures related to uncontrolled environmental factors. A confusing pattern of risk, like that illustrated for the OR_1 estimates in table 1, can also emerge in the presence of J-shaped or U- shaped dose-response trends, due to an excess risk for the two extreme exposure categories. For example, both low and high values of foetal growth were consistently associated with a subsequent risk of developing Neuroblastoma during childhood (7). Another example involves the protective effect consist-

Table 1 - Estimation of the effect of a hypothetical categorical exposure in a simulated case-control study involving 583 cases and 579 controls, using two different reference categories

Exposure categories	Controls	Cases	OR_1	*95% CI*	OR_2	*95% CI*
A	297	310	1.0 (ref.)	-	1.4	*1.1-1.8*
B	168	126	0.72	*0.54-0.95*	1.0 (ref.)	-
C	114	147	1.2	*0.92-1.7*	1.7	*1.2-2.4*

OR_1=Odds Ratio estimates obtained selecting the A category as the referent; OR_2=Odds Ratio estimates obtained selecting the B category as the referent; 95% CI = 95% Confidence Intervals of the Odds Ratio estimates; ref.=reference category ently reported of a low to moderate intake of anti-oxidant and anti-inflammatory chemicals against the risk of developing degenerative diseases; such an effect often tends to disappear or even to reverse when the same compounds are assumed at a either too high or too low concentrations (4). In all these cases, the use of statistical inference via the traditional p-value can help researchers to: a) assess the presence of an association; b) identify a suitable referent category; c) correctly interpret the estimated parameters, and d) formulate adequate biological hypotheses.

It should also be noted that, in the presence of polytomous categorical predictors, the standard method to compute CI is not formally correct. For instance, in the presence of three independent strata a confidence level of 90% (*i.e.*, 95% x 95%) should be adopted in order to obtain a 95% region for the two corresponding point estimates, while in the case of a four-level predictor, the single CI should be calculated at an about 85% confidence (*i.e.*, 95% x 95% x 95%). Furthermore, when strata are obtained from stratification of a continuous variable, they cannot be considered as independent anymore, and the computation of correct 95%CI becomes very cumbersome (10).

In conclusion, we totally agree with the Authors that the abuse of p-value should be strongly discouraged. However, we also agree with Clarice Weinberg's point of view, that "the P-value (and its Bayesian counterparts) has important uses, and should remain an important tool for inference in epidemiology" (11).

Stefano Parodi
Riccardo Haupt
Unit of Epidemiology and Biostatistics,
G. Gaslini Children's Hospital, Genoa, Italy
E-mail: parodistefano@icloud.com

References

1. Consonni D, Bertazzi PA: Health significance and statistical uncertainty. The value of P-value. Med Lav 2017; 108: 327-331
2. 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. doi: 10.1007/s10654-016-0149-3
3. Lash T: The harm done to reproducibility by the culture of null hypothesis significance testing. Am J Epidemiol 2017; 186: 627-635
4. Martucci M, Ostan R, Biondi F, et al: Mediterranean diet and inflammaging within the hormesis paradigm. Nutr Rev 2017; 75: 442-455
5. Nuzzo R: Scientific method: statistical errors. Nature 2014; 506: 150-152. doi: 10.1038/506150a
6. Parodi S, Gennaro V, Ceppi M, et al: Comparison bias and dilution effect in occupational cohort studies. Int J Occup Environ Health 2007; 13: 143-152
7. Rios P, Bailey HD, Orsi L, et al: Risk of neuroblastoma, birth-related characteristics, congenital malformations and perinatal exposures: A pooled analysis of the ESCALE and ESTELLE French studies (SFCE). Int J Cancer 2016; 139: 1936-1948
8. Rothman KJ: Disengaging from statistical significance. Eur J Epidemiol 2016; 31: 443-444. doi:10.1007/s10654-016-0158-2
9. Sterne J, Davey-Smith G: Sifting the evidence - What's wrong with significance tests? BMJ 2001; 322: 226-231
10. Weinberg CR: Invited Commentary: Can Issues With Reproducibility in Science Be Blamed on Hypothesis Testing? Am J Epidemiol 2017; 186: 636-638
11. Weinberg CR: It's time to rehabilitate the P-value. Epidemiology 2001; 12: 288-290