Eur J Epidemiol (2011) 26:899–902
DOI 10.1007/s10654-011-9605-2

METHODS

A novel approach to quantify random error explicitly in epidemiological studies

Imre Janszky · Johan Håkon Bjørngaard · Pål Romundstad · Lars Vatten

Received: 23 March 2011 / Accepted: 29 June 2011 / Published online: 30 July 2011

Abstract The most frequently used methods for handling random error are largely misunderstood or misused by researchers. We propose a simple approach to quantify the amount of random error which does not require solid background in statistics for its proper interpretation. This method may help researchers refrain from oversimplistic interpretations relying on statistical significance.

Keywords Random error · Statistical significance · Confidence intervals · Random error units

Presenting and interpreting random error has been a subject of heated debate ever since the introduction of the *P* value and the concept of statistical significance [1]. Several authors have demonstrated that the concept of statistical significance in many situations is directly misleading and sometimes even harmful [1–6]. Nevertheless, these continuous warnings seem to be largely neglected. The essentially fallacious approach of dichotomizing study results based on whether the *P* value exceeds a prespecified value of 0.05 or not, is still dominating several disciplines, including epidemiology, clinical medicine, psychology and the social sciences. In the majority of scientific journals in these disciplines it is nearly impossible to publish reports that avoid reference to statistical significance.

I. Janszky (✉) · J. H. Bjørngaard · P. Romundstad · L. Vatten
Department of Public Health, Faculty of Medicine, Norwegian University of Science and Technology,
7489 Trondheim, Norway
e-mail: imre.janszky@ntnu.no

I. Janszky
Department of Public Health Sciences,
Karolinska Insitutet, Stockholm, Sweden

A key factor behind the dominance of statistical significance is clearly the lack of knowledge. It has been suggested that the *P* value is probably the most misunderstood statistical concept in research [1, 7]. One of the myths surrounding this issue is that the *P* value is a direct measure of random error or statistical variability. In fact, an essential problem with the *P* value is that it inherently mixes the strength of the association and its precision, thus giving explicit information on neither of them [3, 4, 8]. The strength of the association and its precision are distinct aspects of the data and both have their own essential scientific values. Thus, the use of confidence intervals (CIs) is preferred over the *P* value as it allows the separate assessment of these two distinct phenomena [8]. The point estimate provides information on the observed strength of the association, and the width of the confidence intervals represents random error.

However, it is unfortunate that CIs are also poorly understood and frequently misused [2, 3, 5, 9, 10]. Researchers using confidence intervals are supposed to mentally visualize the underlying *P* value function [4]. However, only a small fraction of researchers is able to do so, and even fewer are likely to practice this mental visualization routinely. Most importantly, far too many researchers do not utilize the rich information provided by CIs, but typically only check whether the 95% CIs contain the null value, i.e., to see whether the results are statistically significant or not [2, 3]. These researchers lose the advantages that CIs can offer and are back to the simple, but flawed approach of dichotomizing study results.

Part of the problem could be that CIs may not be an ideal way to present random error. The absolute width of the confidence intervals for relative measures, such as the odds ratio or the hazard ratio, can be misleading. Theoretically, a study with a confidence interval for an odds

ratio from 0 to 1 has exactly the same imprecision as a study with a confidence interval from 1 to infinity.

Alternative solutions to handle random error have also been suggested, like using Bayesian methodology, likelihood intervals, or presenting the likelihood function[2, 4] but these concepts are at least as complex as that of the confidence intervals and have hardly, if at all penetrated to the research community and become part of common practice.

There could be other, less complex ways to quantify the amount of random error, not requiring a solid background in statistics for their proper interpretation. We propose to present the random error in units analogous to the "meter", i.e., the universally accepted unit of length, which originally referred to the one ten-millionth of the distance from the Earth's equator to the North Pole. Our proposal is to use the amount of the random error present in a hypothetical study as the unit of random error. The proposed hypothetical study is free of any systematic errors and includes one million individuals with an odds ratio of 1 for the association of a dichotomous exposure and the—likewise dichotomous—outcome. To maximize precision half of the study population would be exposed and half would have the outcome of interest. If the amount of random error present in this large, hypothetical "gold standard" study could be looked upon as the unit of random error, then the number of random error units could be calculated in any study using odds ratios for dichotomous exposures or dummy exposure categories by the following simple formula:

$$\text{Number of random error units} = (\text{SE}/0.004)^2$$

The SE is the standard error of the log odds ratio or logistic regression coefficient in the actual study in which we want to assess precision, and provided by all standard statistical outputs. The value 0.004 is the standard error for the log odds ratio in the hypothetical gold standard study, and can be calculated from the well known asymptotic formula:

$$\text{SE} = \sqrt{(1/\text{a} + 1/\text{b} + 1/\text{c} + 1/\text{c})}$$

where a, b, c, and d respectively, refer to those with both the outcome and the exposure, those without the outcome who were exposed, those with the outcome who were not exposed and those without the outcome or exposure, each being equal to 250,000 participants in the proposed gold standard study.

This approach of presenting random error is based on the variance of the log odds ratios or the regression coefficients. The variance of a regression coefficient is a number that is difficult to handle and interpret, and it is seldom reported or used to quantify random error in biomedical studies. Another proposed way to express random error is the confidence limit ratio, i.e., the ratio of the upper to the lower limit of the CI [10]. This is equivalent to the quantity of $e^{3.92*\text{SE}}$ and it allows an order of precision across different confidence intervals to be established. However—and this can be a reason for the relatively infrequent use of this method—it does not offer an explicit quantification of the random error with an easy intuitive interpretation. In contrast, the number of random error units has a simple interpretation. It shows how many times more individuals an actual study would need, providing that the proportion of exposed and those with the outcome will not change, to achieve the precision of the hypothetical gold standard study. For example, consider a study of 100 individuals, half of them exposed to a dichotomous exposure which has no effect on the—likewise dichotomous—outcome, which is also present in half of the individuals. The standard error of the log odds ratio in this study is 0.4 and consequently, the number of random error units is 10,000. If we multiply this study with 10,000 (keeping the proportion of exposed and those with an outcome constant) we arrive at exactly the proposed "gold standard" study. More generally, decreasing the standard error of a study by a factor of $\mathbf{n}$ requires $\mathbf{n^2}$ times as many observations (providing that the distribution of the exposure and outcome is constant). This can be shown by the following:

$$\begin{aligned}\text{SE/n} &= \sqrt{(1/a + 1/b + 1/c + 1/c)} \\ &= \sqrt{(1/(n^2a) + 1/(n^2b) + 1/(n^2c) + 1/(n^2c))}\end{aligned}$$

Our choice of hypothetical standard study was arbitrary. On the one hand, any hypothetical study could serve its purpose as long as the same one is used as standard reference when comparing random error across real-life studies. On the other hand, a study with as little statistical variability as one including 1 million individuals with 50–50% distribution of exposure and outcome might offer some benefits. If we were to choose a small study as a standard, the number of random error units might go below 1 when comparing real life studies to the smaller hypothetical standard. In this case the interpretation of the random error might be awkward as it could imply that a non-integer number of individuals would be needed to achieve the same precision as the standard hypothetical study. Therefore, we would prefer a large study as standard with considerably smaller amount of random error than the great majority of real epidemiological studies. Even epidemiologists involved in register based or multicenter studies can only dream about as precise study as our proposal for the gold standard. This would ensure that 1 REU (random error unit) provides the "atom"—i.e., a "non-dividable" unit—of random error as the number of random error units is not likely to go below 1 and decimal values will not be needed.

In Table 1, we present the number of random error units in several hypothetical studies with different total number of individuals, and different proportions and distributions of the exposure and outcome.

Presenting the number of random error units provides direct and comparable information on the amount of random error in each study. Let us consider three reported odds ratios on the association of fish consumption with gastric cancer risk: 1.4 (95% CI, 0.95–2.0)[11], 0.37 (0.19–0.70)[12] and 2.2 (1.2–3.8)[13]. Just by looking at the confidence intervals from these estimates the amount of random error is not obvious, not even the order of precision between studies is clear for untrained eyes.

The amount of random error in these studies is estimated as 2,446, 7,227 and 5,977 random error units, respectively. Many biomedical researchers and journal editors would probably classify the results with a confidence interval of 0.95–2.0 as "inconclusive". A myth surrounding this issue is that one shall place more trust in a statistically significant estimate than in a non-significant one. As Charles Poole observed: "*confidence intervals are occasionally described as 'wide,' but 'wide' and 'imprecise' often seem nothing more than code words for 'includes the null value' and hence for 'not statistically significant'* "[10]. By presenting the number of random error units it should become clear to everyone that the precision is considerably higher in the first than in the other two studies. Thus, the random error can be low (and therefore the estimates more "trustable") even if a study lacks statistical significance. We believe that the quantification of random error by presenting the random error units may distract attention from whether the intervals contain the null value or not, and we hope this approach could help researchers refrain from using oversimplistic dichotomy in their research.

Moreover, reporting random error units, and explicitly showing the imprecision of a study, could also help to prevent the frequent but pointless discussions about post-hoc power [14]. For example, in a fourth study the odds ratio of the association of fish consumption with gastric cancer risk was 1.0 (0.8–1.3) [15, 16]. The observed, post-hoc power is obviously very low in the study, close to null. In contrast, the precision is rather high as the estimated number of random error units is only 810.

The principles of using units of random error based on gold standard studies can be extended to other measures of association. For example, in the case of using hazard ratios, one can also consider the use of 0.004 in the denominator of the formula. This corresponds to a hypothetical prospective study of one million individuals without censoring, where half of the individuals are exposed to a dichotomous exposure which has no effect on the—likewise dichotomous—outcome, which occurs in half the individuals at the same time during follow up.

Of course, the number of random error units is correct only if the underlying statistical model is correct. Furthermore, it provides no information on systematic errors,

Table 1 Number of random error units in hypothetical studies with dichotomous exposures and outcomes using odds ratios

N of exposed	N of outcome	Total N	Exposure	Outcome		OR (95% CI)	REU
				Yes	No		
500,000	500,000	1 million	Yes	250,000	250,000	1 (0.99–1.01)	1
			No	250,000	250,000		
50	50	100	Yes	25	25	1 (0.46–2.19)	10 000
			No	25	25		
500,000	1,000	1 million	Yes	500	499,500	1 (0.88–1.13)	250
			No	500	499,500		
1,000	500,000	1 million	Yes	500	500	1 (0.88–1.13)	250
			No	499,500	499,500		
1,000	1,000	1 million	Yes	1	999	1 (0.14–7.11)	62 625
			No	999	998,001		
1,000	1,000	10,000	Yes	100	900	1 (0.80–1.24)	772
			No	900	8,100		
1,000	1,000	10,000	Yes	8,169	831	2 (1.67–2.39)	528
			No	831	169		
9,000	1,000	10,000	Yes	831	169	0.5 (0.42–0.60)	528
			No	8,169	831		
1,000	1,000	10,000	Yes	55	945	0.5 (0.38–0.66)	1,276
			No	945	8,055		

REU number of random error units

like biases or confounding—that may often be more important to consider than the random error. As multivariable adjustments influence the precision of the estimates of effect, adjustments will also influence the number of calculated random error units, however, the method of calculating random error units remains the same.

The calculation of the number of random error units is easy and straightforward, it has a simple and intuitive interpretation, and it appears to have some potential advantages. Although it cannot replace CIs, we believe the number of random error units would be a more useful companion to CIs than a *P* value.

References

1. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010;25(4):225–30.
2. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77(2):195–9.
3. Rothman KJ. Curbing type I and type II errors. Eur J Epidemiol. 2010;25(4):223–4.
4. Rothman KJ, Greenland S. Lash TL. Modern epidemiology. Philadelphia: Wolters Kluwer/Lippincott Williams and Wilkins; 2008.
5. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychol Sci. 2004;15(2):119–26.
6. Goodman SN. Toward evidence-based medical statistics. 1: the *P* value fallacy. Ann Intern Med. 1999;130(12):995–1004.
7. Goodman S. A dirty dozen: twelve *P* value misconceptions. Semin Hematol. 2008;45(3):135–40.
8. Rothman KJ. Significance questing. Ann Intern Med. 1986;105 (3):445–7.
9. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Still much to learn about confidence intervals: reply to Rounder and Morey (2005). Psychol Sci. 2005;16(6):494–5.
10. Poole C. Low *P* values or narrow confidence intervals: which are more durable? Epidemiology. 2001;12(3):291–4.
11. Rao DN, Ganesh B, Dinshaw KA, Mohandas KM. A case-control study of stomach cancer in Mumbai, India. Int J Cancer. 2002;99(5):727–31.
12. Pourfarzi F, Whelan A, Kaldor J, Malekzadeh R. The role of diet and other environmental factors in the causation of gastric cancer in Iran–a population based study. Int J Cancer. 2009;125(8): 1953–60.
13. Ward MH, Lopez-Carrillo L. Dietary factors and the risk of gastric cancer in Mexico City. Am J Epidemiol. 1999;149(10): 925–32.
14. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994;121(3):200–6.
15. Buiatti E, Palli D, Bianchi S, Decarli A, Amadori D, Avellini C, Cipriani F, Cocco P, Giacosa A, Lorenzini L, et al. A case-control study of gastric cancer and diet in Italy. III. Risk patterns by histologic type. Int J Cancer. 1991;48(3):369–74.
16. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55: 19–24.