General concepts in biostatistics and clinical epidemiology: Random error and systematic error

Felipe Barraza[a], Marcelo Arancibia[a,b], Eva Madrid[a,b], Cristian Papuzinski[a,b,*]

[a] Cátedra de Metodología de la Investigación Científica, Escuela de Medicina, Universidad de Valparaíso, Valparaíso, Chile
[b] Centro Interdisciplinario de Estudios en Salud (CIESAL), Universidad de Valparaíso, Valparaíso, Chile

*Corresponding author cristian.papuzinski@uv.cl

Citation Barraza F, Arancibia M, Madrid E, Papuzinski C. General concepts in biostatistics and clinical epidemiology: Random error and systematic error. *Medwave* 2019;19(7):e7687

Doi 10.5867/medwave.2019.07.7687

Submission date 27/6/2019
Acceptance date 5/8/2019
Publication date 27/8/2019

Origin This article is one of several "Methodological notes" prepared by the course on Research Methodology of the School of Medicine of the University of Valparaíso

Type of review externally peer-reviewed by four reviewers, double-blind.

Key Words bias, probability, confidence interval, epidemiology, biostatistics

Abstract

Biomedical research, particularly when it involves human beings, is always subjected to sources of error that must be recognized. Systematic error or bias is associated with problems in the methodological design or during the execution phase of a research project. It affects its validity and is qualitatively appraised. On the other hand, random error is related to variations due to chance. It may be quantitatively expressed, but never removed. This review is the first of a methodological series on general concepts in biostatistics and clinical epidemiology developed by the Chair of Scientific Research Methodology at the School of Medicine, University of Valparaíso, Chile. In this article, we address the theoretical concepts of error, its evaluation, and control. Finally, we discuss some current controversies in its conceptualization that are relevant to undergraduate and graduate students of health sciences.

Key ideas

- Error is inherent in biomedical research.
- Systematic error (bias) is associated with weaknesses in methodological design or study execution that can affect the validity of the study results. It can be assessed qualitatively and avoided.
- Random error is the result of variations that occur due to chance and affect the reliability of the investigation. It can be estimated and expressed quantitatively using p-values and confidence intervals. It cannot be eliminated, but it can be controlled by using larger sample sizes and efficient statistical analysis.
- When interpreting research study conclusions, the potential effects of error (both systematic and random) should always be taken into account.

Introduction

Biomedical research, especially when conducted on human beings, is constantly subjected to errors due to the characteristics of its object of study, as well as practical and bioethical limitations. Error assessment is fundamental in the analysis of data, but mainly during the study design, which would allow anticipating the occurrence of systematic errors. Random error, on the other hand, can be expressed quantitatively according to the theory of probabilities, which allows us to estimate the effect of chance on the result of a measurement. Random error can affect the presumed representativeness (reliability) of a sample with respect to the source population, adding uncertainty and imprecision to estimates for population parameters. Given the inherent risk of systematic error and the occurrence of random error, the accuracy or validity of research results cannot be expected or assumed. The accuracy or validity of any measurement process used in research is, on the other hand, a requirement. A valid measurement process is on that is free of bias, where the difference between the estimate and the true value of a population parameter, for example, is low and reliable—that is, reproducible and consistent, or accurate, generating data with little variability among successive measurements[1-3].

Random error (which affects reliability) and systematic error (which affects validity) are two of the main elements evaluated during the development of scientific research and the subsequent critical evaluation by the readers of the published article. Because it is assumed, from a complex point of view, that the studied phenomena are always multicausal and multivariate, considering an association as true and, even more, declaring it as "causal," implies the combination of results from different disciplines and always requires the integration of its context.

This review is the first release of a methodological series of six narrative reviews about general topics in biostatistics and clinical epidemiology. Each article will cover one of six topics based on content from publications available in the main databases of scientific literature and specialized reference texts. In this first review, we analyze different theories and practical elements associated with error in biomedical research, emphasizing its evaluation and control. Finally, we review the current perspectives regarding their theory and some controversies regarding their conceptualization.

Preliminary concepts

A hypothesis is a tentative answer to a research question. In the case of the statistical hypothesis, the construction is based on two assumptions: the null hypothesis (H_0) and the alternative or alternate hypothesis (H_1). H_0 assumes there are no statistically significant differences between specified populations, variables, or other phenomena in the world (and that any apparent differences are due to sampling errors) and therefore inductive inferences (generalizations) about any relationship(s) between them are wrong—that is, the exposure and outcome factors are not related to each other. It is a conservative hypothesis posed in contrast to H_1, the research or work hypothesis, which asserts that observed associations between different phenomena are not explained by chance[4].

The declaration of a null and an alternative hypothesis is essential in inferential statistics, where hypotheses contrast tests are applied to find sufficient evidence to reject the null hypothesis and to support the hypothesis under investigation. However, it should always be kept in mind that the result of a hypothesis test is just one more element for decision making[5] (Example 1)[6].

Example 1. The association between chocolate consumption and cognitive functioning has been studied. In this context, Messerli argues that countries with higher consumption of chocolate have a greater number of Nobel Prize winners because chocolate may be associated with cognitive performance. In this case, the statistical hypothesis of the researcher (H_1) would be that chocolate consumption is correlated with obtaining a Nobel Prize. Therefore, the null hypothesis (H_0) would be that chocolate consumption does not correlate with obtaining a Nobel Prize.

Systematic error (bias)

Systematic error or bias can be understood as the systematic tendency to underestimate or overestimate the estimator of interest because of a deficiency in the design or execution of a study[7]. This bias undermines the study's validity (internal or the degree of agreement between the study results and the true value of the population parameter, and external or the degree to which the results for one study sample can be extrapolated to other populations)[2]. Biases can be associated with any phase of a research study but tend to skew the results in the same direction[2].

Biases that result in overestimation of the magnitude of association between variables are described as positive ("against" the null hypothesis) and biases that reduce the magnitude of an association are described as negative ("in favor" of the null hypothesis). In an extreme case, bias can trigger the inversion of an association, causing, for example, a protective factor to appear as a risk factor; this form of error is called "switch-over bias"[3,8].

When it comes to research with human beings, systematic error is controlled through the study of epidemiology, using the appropriate methodological designs and data collection strategies[2]. There are many different types of bias (https://catalogofbias.org/)[9,10] but they usually fall into three main categories: selection, measurement (or information), and confusion[4]. Selection bias occurs when the relationship between exposure and outcome changes across different groups of study participants (that is, there are systematic differences between the characteristics of the participants)[8] (Example 2).

Example 2. Some research has pointed to the consumption of meat as a risk factor for the development of gastric cancer. To analyze it, a prospective cohort study is designed to compare the rate of five-year survival after diagnosis of gastric cancer between meat consumers (group A) and non-meat consumers (group B). Group A comes from a country with no systematic research on that neoplasm and no specific health system protocol for diagnosis and treatment. Group B comes from a country where routine digestive endoscopies are performed because the health system recognizes a high incidence of gastric cancer in the region. It is concluded that group B has a significantly higher 5-year survival rate. However, it is likely that the longer survival of group B is

due to early diagnosis and treatment rather than to a diet free of meat. In this case, bias was inherent in the selection of the samples because some of their baseline characteristics (e.g., health opportunities) were different. In this case, samples free of selection bias would only have differed in whether they consumed meat or not.

Regarding measurement bias, it has three different forms: bias in the measured phenomenon (for example, memory bias due to differential recall of exposure in a case-control study), bias in the measuring instrument (for example, changes over time in diagnostic criteria), and bias of the observer who makes the measurement[1]. Confusion bias occurs when errors occur in the interpretation of associations between dependent and independent variables due to inadequate control of other variables in the research protocol. The different types of biases will be described in more detail in future articles in this series on the various methodological designs in which they occur most frequently. Due to its complexity and ubiquity, confusion bias will be covered in several reviews in the series.

A confounding variable is one that is associated with both the exposure variable (without being a result of it) and the outcome variable (regardless of its association with the exposure variable) but is not found in the causal path of association[11,12]. This skews or "confuses" the association between exposure and outcome[1]. Confounding variables should not be mistaken for interaction variables that operate as "modifiers of the effect" (those that interact with the exposure variable by modifying the magnitude of its effects on the outcome but are not the cause of the outcome itself)[1].

Researchers should be aware that confusion bias is complex, prominent, and multifactorial[4]. It can be prevented at the study design level (for example, randomization in randomized clinical trials) or controlled during data analysis (stratified analysis or statistical regression models, for example)[4]. Example 3 describes the effect of confusion bias in an observational study conducted in Norway by Strand et al.[12,13].

Example 3. A cohort study conducted to compare 849 children with cerebral palsy to 615,668 children without the pathology concluded that the odds ratio of having a mother with preeclampsia for those with cerebral palsy was 2.5 (95% confidence interval: 2.0 to 3.2). That is, children had a 2.5 times greater chance of having cerebral palsy if their mother suffered from preeclampsia. However, this odds ratio was slightly attenuated (2.1; 95% confidence interval: 1.7 to 2.7) when the association was adjusted for the variable "children who were small for gestational age" during the statistical analysis. Additional adjustments that considered the variable "preterm newborn" reversed the association in favor of preeclampsia, demonstrating that it could be a protective factor for the development of cerebral palsy for children born before 32 weeks who were not small for gestational age (odds ratio: 0.5; 95% confidence interval: 0.5 to 0.8). In this type of study, where the association is observed, the researcher has no control over the exposure variable, so the probability of incurring biases is greater. The authors controlled for confusion bias during statistical analysis using a statistical regression model known as logistic regression, which is often used to identify and evaluate confounding variables that might not emerge otherwise.

Random error (chance)

Random error is associated with variations resulting from chance that are inherent in all research and cannot be eliminated; this type of error can therefore influence results, even when biases have been properly controlled[7], and compromise the reliability of the investigation. Three main factors are associated with random error in study results[14]: the degree of individual and inter-individual variability, the sample size, and the magnitude of the differences (with the likelihood of it being caused by chance falling as the difference found in the comparison increases).

Observations that deviate from the true value of a variable in any sense are attributed to random error[2]. Random error is unpredictable but can be reduced by using larger sample sizes and efficient statistical analysis[14]. This reduction implies that statistics control random error[2], and that probability is related to the chance occurrence[7]. Therefore, adequate estimation of the sample size should counteract the effect of chance in the study. It should be noted, however, that a study's sample size would not be an indicator of its internal validity. In other words, sample size is not directly associated with the level of bias of a research investigation[14]. Estimation of random error is carried out through two procedures: hypothesis contrast tests (p-value) and confidence intervals[15].

P-value

The value of p (probability) is the likelihood of observing an apparent outcome, assuming that the null hypothesis is true. That is, the p-value is the probability of chance occurrence in the case that the null hypothesis is true (assuming that the phenomena under study are not related)[4]. The p-value answers the question "What is the probability of finding that association just by chance?" It is also a measure of the strength of the evidence against the null hypothesis[14], as it can be understood as the probability of obtaining certain results given that the hypothesis that the researcher "wants" to reject is fulfilled. Therefore, intuitively, if the p-value is very small, the null hypothesis is rejected, and the research hypothesis is achieved[16].

Traditionally, the value of p is 0.05 (that is, when applying statistical methods that analyze the probability of occurrence, the associations have happened due to chance no more than 1 in 20 times, or 5% or less of the times)[4]; this suggests a probability that the null hypothesis will be rejected by mistake 5% of the time. However, there is no scientific reasoning behind the value 0.05 in itself; it is an arbitrary convention[17]. This cutoff point is referred to as "statistical significance" (the value at which the null hypothesis can be rejected). This does not necessarily indicate that something important ("significant") has happened but should be interpreted as a calculation showing that something "meaningful" has happened[4,18]. Some current initiatives have proposed lowering the threshold of the level of significance from 0.05 to 0.005[19].

Different hypothesis tests are linked to different p-values; the proper choice of p-value depends on the study design and random variables.

All are a function of the difference between the values observed in the study and those that would be observed if the null hypothesis were true, given the variability of the sample[15]. Another way of representing p-values is as fractions whose denominators (variability of the result) decreases as the sample size increases and numerators increase when the difference between the observed values and the expected values is greater[14].

Based on the information above, there are two types of errors associated with chance. The first is the type I error, conceptualized as the probability of rejecting H_0 given that H_0 is true. This occurs when a study outcome suggests an association between variables that does not really exist. Thus, the statistical significance mentioned above constitutes the limit of type I error, whose numerical value is called α[20]. This type of error is found most frequently in clinical studies that seek to analyze a large number of associations simultaneously. Examples include a cohort study that analyzes multiple variables for the same exposure, a clinical trial with different subgroup analyses, and a case-control study that explores countless risk factors together[20].

On the other hand, when there is an association, but the difference is not investigated by the study, a type II error occurs. This represents the probability of not rejecting H_0, since H_0 is false. The symbol for a type II error is β. The complement of β (1-β) corresponds to the statistical power of the study (the probability of finding a difference, if it exists, in other words, verifying the research hypothesis). The power of a study is usually 0.8 to 0.9, meaning that it is 80% to 90% likely to detect the proposed difference, and that the result has statistical significance[20]. Example 4 shows the interpretation of the p-value in a study of biomarkers in severe mental disorders[21].

Example 4. Several investigations have reported an increased level of proinflammatory cytokines in people with severe psychotic and affective disorders and those who have suffered psychological trauma during childhood. To investigate this further, markers of inflammation and a history of childhood trauma were studied in people with schizophrenia, schizoaffective disorder, bipolar disorder with psychotic phases, and healthy people. The authors concluded that people with schizophrenic disorders had significantly higher levels of interleukin-6, tumor necrosis factor α, and C-reactive protein (proinflammatory cytokines) compared to healthy people, as well as significantly higher levels of tumor necrosis factor α compared to people with bipolar disorders (with a p-value < 0.05 for all comparisons). It was also found that exposure to childhood sexual abuse had a significant association ($p = 0.018$) with C-reactive protein levels in people with schizophrenia. In this study, the authors established a level of significance of 5% (0.05). That is, if an association has a p-value lower than this threshold when applying a statistical test, the null hypothesis (H_0) can be rejected in favor of the alternative (research) hypothesis (H_1), which, in this case, was that inflammatory markers are associated with severe mental disorders and the presence of childhood trauma. Thus, significant associations between the parameters studied were found. Specifically, it was determined, with a 95% probability, that the observed associations were not due to chance but were explained by a different underlying mechanism.

The question of whether the p-value depends on the sample size, with small samples more vulnerable to random error than larger ones, has been the subject of extensive debate. This is the reason for some of the strongest criticism of the use of hypothesis tests (the idea that the rejection of a hypothesis depends on sample size[22], since the study conclusions will be limited if only a small portion of the population is evaluated, but the rejection of the null hypothesis would be virtually assured if a large part is evaluated)[22]. However, some authors oppose this criticism, claiming that, when studying discrepancies in results of randomized clinical trials with large and small sample sizes, the differences found are not explained by sample size but by the control of biases, especially confusion bias. For this reason, the general rule would be mostly concordance between results and not the differences[14,23]. Due to the limited amount of information that the p-value can provide on its own, another way to quantify chance is using confidence intervals.

Confidence intervals

Confidence intervals consist of a range of values in which the real value of the parameter can be determined with a certain probability[7]. Therefore, confidence intervals reflect the degree of uncertainty. As already mentioned, a larger sample will result in a more precise confidence interval for estimation of a population parameter (that is, a narrower range of values, indicating a lower effect of chance on the estimate). Like the estimation of the p-value, the estimation of the confidence intervals requires statistical inference, since a critical value in the interval that indicates the lack of association between two variables is excluded. In the case of indicators whose formula is a quotient (for example, relative risk and odds ratio), this value is 1, and when the indicators' formula is derived by subtracting the risk of one group from the risk of another (for example, absolute risk reduction), the value is 0. In both cases, the values represent the points at which an event is equally likely in both groups[18]. Values that exceed the limits of the confidence interval may not always be entirely excluded, but it would be reasonable to think that it is highly unlikely to find the actual value of the parameter beyond these limits[24]. Example 5 shows the interpretation of the confidence interval based on the results of the study by Strand et al. cited in Example 3[13].

Example 5. Initially, researchers found that having a mother with preeclampsia increased the chance of having cerebral palsy by 2.5, assuming that this association was significant, based on its 95% confidence interval of 2.0 to 3.2. In other words, the confidence interval did not include the value 1, indicating that there was no association between the study variables. This detail is important, since the odds ratio is calculated using a quotient. Subsequent statistical analyses showed that preeclampsia would be a protective factor against cerebral palsy in children born before 32 weeks who were not small for gestational age, since the odds ratio was 0.5, with a 95% confidence interval of 0.5 to 0.8. The interpretation is therefore the same: with a probability of 95%, the association between mothers with preeclampsia and children without cerebral palsy is explained by an underlying mechanism other than chance.

Perspectives and final considerations

Considering an association as "true," in other words, not explained by bias or chance, implies thinking about causality[25], integrating what is known thus far about the mechanism underlying the phenomenon under study. This encourages us to use new approaches rather than interpreting statistical associations superficially, which entails more complexity in our thinking, given various problems pointed out below.

In order to extrapolate a result found in a sample at the population level (that is, generalize based on a certain outcome), the conclusions must not only be based on a statistical procedure or on the level of representativeness of the sample with respect to the population but must also incorporate existing knowledge about the phenomenon under study[26]. Causality should be studied considering the previous findings of other studies in the field. However, the possibility of integrating these findings into the analysis itself cannot be accurately measured.

The analytical process associated with statistical inference through hypothesis tests excludes some important factors, such as biological plausibility and the body of existing evidence. Different approaches have emerged as alternatives to this process, including Bayesian methods[27]. Bayesian methods integrate previous experiences in the inferential process, since it is assumed that cumulative experience applying a certain hypothesis can and should contribute to its verification[16]. Here, the researcher expresses aprioristic points of view probabilistically, and these are added to the formal data analysis[28].

Although the Bayesian approach is not yet widely used in biomedical research[16], there is extensive research on it, with positive results. In the meantime, even though this approach has the same theoretical framework as the p-value (the frequentist approach to probability), scientific publications have promoted use of the p-value, with confidence intervals for at least three decades[29,30]. Confidence intervals are based on the same statistical framework as p-values but provide more information about difference between outcomes and chance in the measurement process[22].

Many scientific articles have focused on the p-value, which, as mentioned above, is a quantitative mechanism for assessing chance. According to some authors, even scientific research has focused on chance[4]. This has occurred at the expense of critical evaluation of biases, whose assessment is qualitative. Given a theory that, on the one hand, sheds light on which methodological designs are conducive to each type of bias, and, on the other hand, interprets the study findings in the context of what is already known about the phenomenon, it is worth asking in what sense and to what extent could bias have affected these results? Can I believe what I see? That is, the results are not reliable on their own; their value depends on the accuracy of the measuring processes that derived them.

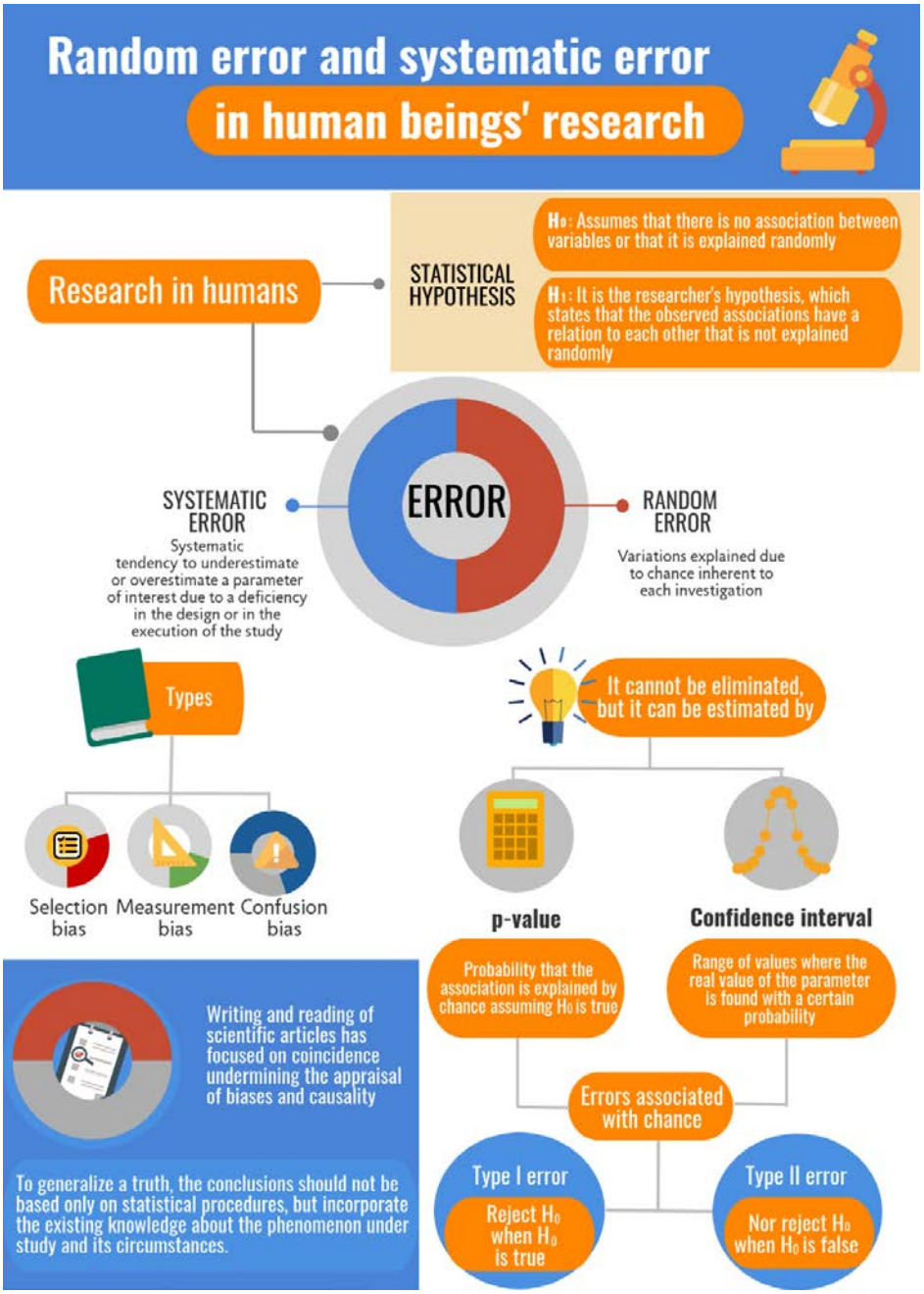
Notes

Funding
None.

Competing interests
None.

Language of submission
Spanish. The English translation of the originally submitted article has been copyediting by the Journal.

Figure 1. Diagram of random error and systematic error.



References

1. Ruiz J. Variables y fuentes de error en la medición. Confusión, sesgo y azar. In: Ruiz Á, Gómez-Restrepo C, editors. Epidemiología clínica: investigación clínica aplicada. 2nd ed. Bogotá: Editorial Médica Panamericana; 2015. | Link |
2. Martínez-González M, Toledo J, López-Fidalgo J. Intervalos de confianza y contraste de hipótesis. In: Martínez M, Sánchez-Villegas A, Toledo E, Faulin J, editors. Bioestadística amigable. 3rd ed. Barcelona: Elsevier; 2014. | Link |
3. Delgado-Rodríguez M, Llorca J. Bias. J Epidemiol Community Health. 2004 Aug;58(8):635-41.| PubMed |
4. Ghaemi S. Hypothesis-testing: the dreaded p-value and statistical significance. In: A clinician's guide to statistics and epidemiology in mental health: measuring truth and uncertainty. Cambridge: Cambridge University Press; 2009.
5. Fisher R. The design of experiments. 9th ed. New York: MacMillan Pub Co; 1971. | Link |
6. Messerli FH. Chocolate consumption, cognitive function, and Nobel laureates. N Engl J Med. 2012 Oct 18;367(16):1562-4. | CrossRef | PubMed |
7. Coggon D, Geoffrey R, Barker D. Measurement error and bias. In: Epidemiology for the uninitiated. 5th ed. London: BMJ Books; 2003.
8. Araujo M. General validity criteria in clinical trials. Medwave 2012 Mar/Abr;12(3):e5338 . | CrossRef |
9. Araujo M. Some biases and limitations present in biomedical literature: Part 1. Medwave 2013;13(3):e5647. | CrossRef |
10. Araujo M. Some biases and limitations present in biomedical literature: Part 2. Medwave 2013;13(4):e5668. | CrossRef |
11. de Irala J, Martínez-González MA, Guillén Grima F. [What is a confounding variable?]. Med Clin (Barc). 2001 Oct 6;117(10):377-85. | PubMed |
12. Ananth CV, Schisterman EF. Confounding, causality, and confusion: the role of intermediate variables in interpreting observational studies in obstetrics. Am J Obstet Gynecol. 2017 Aug;217(2):167-175. | CrossRef | PubMed |
13. Strand KM, Heimstad R, Iversen AC, Austgulen R, Lydersen S, Andersen GL, et al. Mediators of the association between pre-eclampsia and cerebral palsy: population based cohort study. BMJ. 2013 Jul 9;347:f4089. | CrossRef | PubMed |
14. Araujo M. Random error in clinical trials. Medwave 2012 Jun;12(5):e5423. | CrossRef |
15. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016 Apr;31(4):337-50. | CrossRef | PubMed |
16. Jiménez-Paneque R. The questioned p value: clinical, practical and statistical significance. Medwave 2016 Sep;16(8):e6534. | CrossRef | PubMed |
17. Salsburg D. The lady tasting tea: how statistics revolutionized science in the twentieth century. New York: MacMillan Pub Co; 2001. | Link |
18. Madrid E, Martínez F. Statistics for the faint of heart - how to interpret confidence intervals and p values. Medwave 2014;14(1):5892. | CrossRef |
19. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA. 2018 Apr 10;319(14):1429-1430. | CrossRef | PubMed |
20. Araujo M. Random error in clinical trials (II). Medwave 2012 Jul;12(6):e5437. | CrossRef |
21. Quidé Y, Bortolasci CC, Spolding B, Kidnapillai S, Watkeys OJ, Cohen-Woods S, et al. Association between childhood trauma exposure and pro-inflammatory cytokines in schizophrenia and bipolar-I disorder. Psychol Med. 2018 Dec 18:1-9. | CrossRef | PubMed |
22. Benavides A, Silva L. Contra la sumisión estadística: un apunte sobre las pruebas de significación. Metas Enferm. 2000;3(27):35-40.
23. Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA. 1996 Oct 23-30;276(16):1332-8. | PubMed |
24. du Prel JB, Hommel G, Röhrig B, Blettner M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2009 May;106(19):335-9. | CrossRef | PubMed |
25. Doll R. Proof of causality: deduction from epidemiological observation. Perspect Biol Med. 2002 Fall;45(4):499-515. | PubMed |
26. Cohen J. The earth is round (p < .05). Am Psychol. 1994;49(12):997-1003.
27. Silva L. El advenimiento de la era bayesiana. In: Cultura estadística e investigación científica en el campo de la salud: una mirada crítica [Internet]. Madrid: Ediciones Díaz de Santos; 1997. | Link |
28. Silva L, Benavides A. Apuntes sobre subjetividad y estadística en la investigación en salud. Rev Cuba Salud Pública. 2003;29(2):170-3. | Link |
29. Evans SJ, Mills P, Dawson J. The end of the p value? Br Heart J. 1988 Sep;60(3):177-80. | PubMed |
30. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986 Mar 15;292(6522):746–50. | Link |

Correspondence to
Escuela de Medicina, Universidad de Valparaíso
Angamos 655, Oficina 110
Reñaca, Viña del Mar
Chile.