Scand J Work Environ Health 1997;23(3):232-235

<https://doi.org/10.5271/sjweh.204>

Issue date: Jun 1997

Statistical significance -- a misconstrued notion in medical research

by [Nurminen M](#)

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/9243735

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Statistical significance — a misconstrued notion in medical research¹

by Markku Nurminen, PhD²

Nurminen M. Statistical significance — a misconstrued notion in medical research. *Scand J Work Environ Health* 1997;23(3):232—5.

The P-value is the significance probability of obtaining a value of the test statistic that is as extreme, in relation to the null hypothesis, as that observed. Medical researchers may, in some situations, disagree on its appropriate use or on its interpretation as a summary measure of consistency with the null hypothesis in a particular data set. More informative statistical measures such as the likelihood ratio and the Bayesian posterior probability have been suggested for drawing inferences from clinical trials and epidemiologic studies. Causal inference is not statistical in nature; rather it strives to provide scientific explanations or criticisms of proposed explanations that would describe the observed data pattern. In this context, it is important to remember that a finding may not be medically important, or a causal hypothesis may even not be true even if a study shows a significant P-value.

“Statistically significant” is a chronically misinterpreted concept in clinical trials and epidemiology. The misconception can be caused by both the confusing terminology and the difficult theory of statistics. The standard language word “significant” has a special meaning in statistical research: the consistency of data with a hypothesis is measured by the “significance probability” or the P-value. Finney (1) has proposed that one should always add the adverb “statistically” in conjunction with the word “significant” whenever its meaning could otherwise be in doubt.

Significance testing prevails as a general method of analysis in medical research although the overemphasis on the use of the P-value has long been criticized (2). Researchers obviously believe that it is not worthwhile to submit a manuscript for publication to a journal unless it contains a significant P-value. Significance testing is an apparently objective way to decide whether a so-called null hypothesis³ (eg, treatment A is as good as treatment B) remains valid or should be rejected and the study hypothesis be accepted in its place (eg, treatment A is better than treatment B). Instead of the P-value, computations of more informative statistical measures have been suggested. Such statistics include the P-value function (3), which yields the significance of also other hypotheses and not merely that of the null hypothesis,

and the likelihood ratio test (4), which compares 2 rival hypotheses. Some scientific journals (eg, *Cancer Research*) instruct the authors to indicate the significance of their findings using an appropriate statistical analysis. Other journals, such as the *British Medical Journal* (5) and *The Lancet* (6), have recommended that significance testing be replaced by the computation of a confidence interval. Certain statisticians reject significance testing categorically. (See, eg, reference 7.) Respected epidemiologists like Rothman (8) and Greenland (9) would not ban significance tests, but they hold the view that the tests appear to have produced much more harm than good in social and health sciences.

The traditional (“frequentist”) Neyman-Pearson school of statistics and the alternative Bayes school interpret the notion of probability underlying the test of significance differently. The frequentist statisticians define the P-value as the probability of the observed outcome in a study plus the probabilities of the more extreme (unobserved) outcomes — that is, as a relative frequency, or proportion, in large samples — assuming that the observations are generated according to a given probability model. The P-value measures whether a null hypothesis is compatible with the data or not. It is, however, totally contrary to the spirit of significance testing to compare the P-value with preset levels, which are

¹ This commentary was published in Finnish as an editorial in *Duodecim* vol 113, no 4, 1997.

² Department of Epidemiology and Biometry, Finnish Institute of Occupational Health, Helsinki, Finland.

³ Null hypothesis is an exact statistical formulation for the studied assumption (hypothesis) to be incorrect: for example, the difference of 2 groups' mean values equals 0. Assuming that the null hypothesis prevails, one can make deductive inferences about the correctness of the study hypothesis, which is often formulated in less exact terms.

Reprint requests to: Dr M Nurminen, Finnish Institute of Occupational Health, Topeliuksenkatu 41 a A, FIN-00250 Helsinki, Finland.

conventionally chosen as 5, 1 or 0.1%, and to interpret the result in a rigidly different manner depending on whether the P-value is below or above a certain level. [These reference levels are often marked with 1, 2 or 3 asterisks (*), but they do not need to be considered in the same light as the stars indicating the quality of hotels.] Significance testing is not to be regarded as decision-making but as statistical inference. Occasionally one sees the frequentist P-value being interpreted as giving the probability for the statement that "the null hypothesis is true" or that "the result is a random finding". The former interpretation is surely wrong because the computation of the P-value explicitly assumes that the null hypothesis is true. The latter interpretation is problematic since, in a frequentist analysis, one can never infer definitely whether a single hypothesis pertaining to the considered parameter⁴ (eg, the difference between mean values) is true or not or whether the unknown value of the studied parameter lies within, say, the 95% confidence interval computed from a particular experimental material. The frequentist statisticians can only state that, if the experiment were repeated sufficiently many times, then approximately 95% of the computed intervals (which are stochastic variates) would cover the true value of the studied parameter (which is a constant of nature).

In the interpretation of the P-value one must also consider the amount of information contained in the data (the "power" of a test). Miettinen (10) provides the following guidelines for interpretation: (i) if information is very sparse, one should not analyze the data at all; (ii) if information is very ample, the P-value is too sensitive to be useful and, instead of testing, one should estimate the magnitude of the effect; (iii) if the amount of information is neither very sparse nor very ample, one may infer that (a) a very small P-value supports the study hypothesis, (b) a small P-value does not discriminate between the study hypothesis and the null hypothesis, and (c) a moderately or especially large P-value is relatively less consistent with the study hypothesis than with the null hypothesis, which speaks for the refutation of the study hypothesis.

A Bayesian statistician overcomes the interpretative problems of significance testing by viewing probability as a degree of personal belief of the correctness of a study hypothesis. This subjective probability is based on investigative foreknowledge regarding the uncertainty of the study hypothesis and the preconception which one has of it and which will be modified via a model as new empirical evidence accrues. Bayesian statistical theory produces a "posterior" probability distribution of the

studied hypothesis, by means of which one can inductively state, for instance, that "with a 95% probability treatment A is more effective than treatment B at least in 10% of the cases and at most in 20% of the cases." Different experts will often have different preconceptions of the credibility of the studied hypothesis, but in a Bayesian analysis these prior beliefs can naturally be reflected by the presentation of several prior distributions in the context of the same data (11).

The Bayesian approach also avoids the frequentist problem related to the testing of multiple hypotheses in 1 data set or the simultaneous testing of a single hypothesis in many subsets of data. For example, when one studies the differential diagnosis of malignant mesothelioma and lung carcinoma with the aid of genetic alterations, one can examine 10's of different chromosome changes. Frequentist statisticians try to control the occurrence of false significant findings by applying more strict levels of significance. By using this procedure, for example, a significant difference ($P = 0.004$) of the frequency of changes observed in chromosome 22 between patient groups becomes nonsignificant if one accounts for the respective tests made for the chromosomes 1,...,21 in the same investigation and corrects the 5% critical level to $0.05/21 = 0.0024$. According to the Bayesian way of thinking there is no reason to correct a particular P-value merely because other variables were also considered in the same study. The Bayesian solution of the problem is to define the prior joint likelihood of the mutually dependent hypotheses, which would appear to be a scientifically more rational procedure than a mechanistic correction of the P-value. The specification of the prior likelihood function is, however, a challenging data-analytic task, especially in problems involving many parameters (12).

Frequentist inference is thus problematic. Why isn't everyone then a Bayesian (13)? The answer is dictated by practice. For example, the Bayesian likelihood ratio test is harder to compute than the frequentist significance test. Ten years ago, Bayesian analytic solutions of even the simplest epidemiologic problems were difficult to tackle (14). Nowadays, however, simulation modeling techniques make the performance of a Bayesian analysis possible also in more complex biomedical applications (15), in which the frequentist and Bayesian analyses do not necessarily result in the same inferences (16). This being the case, the Bayesian methods will inescapably be used in clinical medicine and epidemiology (17). During the transition period, medical scientists should prepare for the change by familiarizing themselves with the Bayes-

4 A parameter is a quantity which partly or fully determines a probability distribution. A parameter is not directly measurable, but, using a distribution model, one can describe the kind of samples associated with particular values of the parameter. Considering the compatibility of the data with the model, one can estimate the most likely value of the parameter.

sian methods (18). The natural simplicity of the Bayesian concepts is appealing.

The role of statistics in cause-effect studies depends on the study design. The traditional theory of statistics was created for randomized experiments. Thus in clinical trials, in which the treatment of patients is randomized, the results produced by customary analysis (the P-value, the confidence interval of a parameter, the likelihood ratio) are interpretable quantities from the point of view of causal inference. In nonexperimental (eg, epidemiologic) studies, in which the exposure of persons is not randomized, probabilistic interpretations of conventional statistics are not necessarily justified and can lead to incorrect inferences of nonrandomized studies (19). Can thus, for example, the P-value be interpreted in any reasonable way in nonrandomized studies? As a remedy to this problem, Greenland (20) suggests that, in the data analysis, one should separate the following aspects from each other (i) the description of the data variability by means of graphic displays or simple summaries, (ii) the profiling of distributions or relations being sought from the data in comparisons with statistical models (pattern recognition, data smoothing), and (iii) scientific inference. Greenland (20) contends that statistical analysis is limited to stage 2, in which a statistical measure, "such as a P-value is not a data summary; rather it is a convolution of the data with some model or preconceived notion about the process that generated the data [p 227]". One should use modern techniques of statistical analysis to examine the impact of discrepant observations on the outcome measures (influence analysis) and the effects of departures from model assumptions on the stability of the findings (sensitivity analysis). Causal inference is not statistical by nature; rather it strives to (i) determine scientific explanations that would explain the results of statistical analyses in a logically coherent way and (ii) criticize proposed explanations that would not lead to the observed data pattern (20).

"Clinical significance" is determined in population studies, for example, as the magnitude of the difference in the mean values between the experimental group and the comparison group. In large population groups even a small difference becomes statistically significant, whereas in small samples a clinically significant observation can remain statistically nonsignificant. Two recent cohort studies on reproductive health furnish examples of surveys in which the size of the material was a significant factor. The notable sample size (over 5000 people) of a Danish study (21) permitted the expression of minimal differences, whereas, in an American study (22), the small number of exposed persons (only 27) prevented the presentation of differences that were not big.

On the other hand, although the difference would be small on the group level, a clinical finding can be of decisive importance for some persons who belong to a

risk group. In a Finnish epidemiologic study (23), the risk of dying of coronary disease in a cohort of 343 industrial workers exposed to carbon disulfide was over 2-fold relative to the risk of a same-sized, individually matched comparison cohort. The researchers discussed several biochemical mechanisms that would explain why carbon disulfide exposure caused the increased risk of coronary mortality. A possible indirect mechanism might have been high blood pressure. On a group level, the differences in the mean values of the subjects' blood pressures were statistically significant, although the differences were relatively small [difference in systolic pressure 8 mm Hg (1.1 kPa) and in diastolic pressure 3.5 mm Hg (0.5 kPa)]. If a worker had, in addition, other risk factors, even a minor elevation of blood pressure could be a danger. The researchers estimated that high blood pressure was a causal factor in every 6th death due to coronary heart disease, which was originally caused by carbon disulfide exposure (23).

It is not very remarkable if a large study produces a statistically significant result. The finding can be medically important only if one's colleagues still believe in the result after having read the discussion of its significance without reference to P-values.

Acknowledgments

I thank Tuula Nurminen for her valuable comments.

References

1. Finney DJ. On biometric language and its uses. *Biom Bull* 1994;11:2-4.
2. Yates F. The influence of "Statistical Methods for Research Workers" on the development of the science of statistics. *J Am Stat Assoc* 1951;46:19-34.
3. Folks JL. Ideas of statistics. New York (NY): Wiley, 1978.
4. Cox DR. The role of significance tests. *Scand J Stat* 1977;4:49-70.
5. Langman MJS. Towards estimation and confidence limits [editorial]. *BMJ* 1986;292:716.
6. Lancet. Report with confidence [editorial]. *Lancet* 1987;1(8531):488.
7. Oakes M. Statistical Inference. Chesnut Hill (MA): Epidemiology Resources Inc, 1990.
8. Rothman KJ. Modern epidemiology. Boston (MA): Little, Brown and company, 1986.
9. Greenland S. Foreword. In: Oakes M. Statistical inference. Chesnut Hill (MA): Epidemiology Resources Inc, 1990: vii-viii.
10. Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. Albany (NY): Delmar Publishers Inc, 1985.
11. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;313:603-7.
12. Dixon DO, Simon, R. Bayesian subset analysis in a colorectal cancer clinical trial. *Stat Med* 1992;11:13-22.
13. Efron B. Why isn't everyone a Bayesian (with discussion)? *Am Stat* 1986;40:1-11.

14. Nurminen M, Mutanen P. Exact Bayesian analysis of two proportions. *Scand J Stat* 1987;14:67—77.
15. Gilks WR, Clayton DG, Spiegelhalter DJ, Best NG, McNeil AJ, Sharples ED, et al. Modeling complexity: applications of Gibbs sampling in medicine (with discussion). *J R Stat Soc [Serie B]* 1993;55:39—52.
16. Diaconis P, Freedman D. On the consistency of Bayes estimate (with discussion). *Ann Math Stat* 1986;14:1—67.
17. Freedman L. Bayesian statistical methods: a natural way to assess clinical evidence [editorial]. *BMJ* 1996;313:569—70.
18. Berry DA. *Statistics, a Bayesian perspective*. Belmont (CA): Duxbury Press, 1996.
19. Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990b;1:421—9.
20. Greenland S. Summarization, smoothing, and inference in epidemiologic analysis. *Scand J Soc Med* 1993;21:227—32.
21. Hedegaad M, Henriksen TB, Secher NJ, Hatch MC, Sabroe S. Do stressful life events affect duration of gestation and risk of preterm delivery? *Epidemiology* 1996;7:339—45.
22. Landbergis PA, Hatch MC. Psychological work stress and pregnancy-induced hypertension. *Epidemiology* 1996;7:346—51.
23. Nurminen M, Mutanen P, Hernberg S, Tolonen M. Quantitated effects of carbon disulfide exposure, elevated blood pressure and aging on coronary mortality. *Am J Epidemiol* 1982;115:107—18.