REVIEW

Review on significance debate

C. SARDU, P. CONTU
Department of Public Health, University of Cagliari, Italy

Key words

Inference • P value • Significance

Introduction

The terms "p value" and "significance" are probably the most used statistical concepts in science literature [1]. In the field of epidemiology, the key role of significance is clearly visible in all kinds of epidemiological studies: experimental (comparison between two drugs in a clinical trial), analytic (case control or cohort studies designed to analyse health effects of a specific exposure), and descriptive (distribution of a disease by age, sex, or geographic location). Finding out if an association is real or due to chance, relevant or not, caused by confounders or influenced by interactions, is the common task for all these types of research, while correctly interpreting the p value is the shared problem.
A series of articles recently published confirmed our personal experience emphasizing significance interpretation as a topical point in epidemiology [2-7]. As states Smith "the debate about the appropriate place of p values in scientific inference is still alive" [2].
Specifically, the controversy about the correct interpretation of p value has arisen since this parameter was proposed in 1925, and there is still strong disagreement and confusion about this [8].
This paper aims to discuss the main issues related to significance. The article is structured as a glossary; it is not merely descriptive, including not only "technical terms" (power, type I error, confidence interval), but also a number of "interpretative terms" in order to illustrate the debate about significance since its origin, including our personal view.
This glossary is intended mainly for health professionals, who have to deal with significance, in order to identify relevant health determinants, to develop health promotion programmes, risk assessment, to realize risk communication programs, to plan vaccination campaigns, and to evaluate the effectiveness of their actions.

Statistical Inference

There is no complete consensus in the definition of the main function of statistical inference. Inference literally means the process of forming an opinion based on what you already know.
Statistical inference is the process of drawing conclusions about a population, based on the analysis of a sample. "By means of statistical inference it is possible to draw conclusions about models which potentially may have generated the data, and to apply the results to the entire population" [9]. The generalisation of the results from the sample to the population is burdened by some level of uncertainty. According to some Authors the objective measurement of this uncertainty, and so the p value is one of the main goals of statistical inference [9]. Other Authors emphasize the primary function of statistics in providing clear and precise description, and consider the calculation of p value as the least important role of a statistical inference [10].

Frequentist School of Inference

The Frequentist School methodology is based on the assumption that an experiment is repeated an infinite number of times. The evaluation of the inferential process should be made using these virtual repetitions of experiments [9]. In other words it has been demonstrated that if all possible samples from a population were selected, and the data analysis was repeated for each sample, most samples would give similar results (common value), but some, just by chance, would give very different results (rare value). So, according to the frequentist school, the result from one sample has to be compared to the other potential results indicating how rare it is.
Critics state that inference must be based only on results from real experiments, and consequently using methods based on virtual results is incorrect [9].
In spite of this widespread criticism the frequentist school is the most used method of inference, especially the techniques of Hypothesis testing and Confidence Interval [9, 11].

Hypothesis Testing

The procedure of hypothesis testing assesses the uncertainty linked to a parameter calculated from a sample, verifying a null and an alternative hypothesis [12].
The null hypothesis is the statement that no difference exists between groups, or that there is no association between a determinant and an outcome. If the null hy-

pothesis is true then the difference (or association) seen in a study results from chance [13, 14].
The alternative hypothesis is in contradiction to the null hypothesis, stating that some difference between groups exists, or that there is some association between a determinant and an outcome [15].
The hypothesis testing procedure evaluates, by means of statistical tests, if the observed data are more consistent with the null or the alternative hypothesis, calculating how likely the observed difference or association could arise by chance; if this probability is low the null hypothesis should be rejected.

Statistical tests

Statistical tests provide a systematic approach to decide if the null hypothesis should be rejected or accepted, and to assess uncertainty linked to this decision.
A test value is calculated from sample data with appropriate statistical procedures (according to the characteristic of examined data).
A large test value indicates a departure from the null hypothesis. To evaluate if the obtained test value is large enough to reject the null hypothesis, it is assumed that the calculation is repeated an infinite number of times with different random samples drawn from the same population. If the test value is larger than most of the values obtained from the different random samples, this could be uneasily attributed to chance, and so the null hypothesis should be rejected.
This methodology does not address the issue of whether an association is causal or not: the concept of association refers to a dependence, which may or may not be casual, between two or more variables [16, 17].

p-value

The methodology of p value was first proposed as part of an inference method by Fisher in 1925, and later in 1933 it was introduced again in the context of hypothesis testing by Neyman and Pearson [1].
The p value (short for probability value) is defined as the probability of obtaining by chance, from the observed data, the observed test value or a larger one. The p value is considered as summarizing the statistical evidence of an experiment, and it is interpreted as discrediting the null hypothesis if p is small and in favour of the null hypothesis if p is large [1].
The p value is influenced by both the magnitude of the effect and the sample size, and if these factors are not taken into account, two opposite consequences are possible [18].
On the one hand, if the selected sample is too large, the experiment might result in a small p value even though the magnitude of the effect is not relevant.
On the other hand if the sample is too small, the study may not result in a small p value even if the magnitude of the effect is really relevant.

Type I and Type II Error

When the procedure of hypothesis testing is performed, the probability of making a wrong decision exists, and two types of errors are possible [12]:
– type I (α error) that is the probability of rejecting the null hypothesis when it is true; in a clinical context it is the probability of "false positive": that a healthy person is erroneously considered sick;
– type II (β error) that is the probability of accepting the null hypothesis when the alternative is true; in a clinical context it is the probability of false "negative": that a sick person is erroneously considered healthy.

The concept of α error, or false positive, is linked to the term specificity. The specificity is the ability of a clinical test to correctly individualize the absence of a disease when it is really absent, consequently when a test is not specific it also indicates the disease in a healthy subject.
In designing a study both types of error should ideally be minimized [19]. Nevertheless researchers usually pay more attention to type I error than to type II, without considering that the decrease in the probability of α error is often achieved with an increase in the probability of β error [19].

Level of a Test

The probability of making a type I error is called the level of a test or the level of significance [20]. In other words the level of a test is the probability of rejecting the null hypothesis when in fact it is true. According to the original approach of hypothesis testing proposed by Neyman-Pearson a researcher should decide in advance the level of α [12]. Then he should select and perform the appropriate statistical test. The decision to reject or accept the null hypothesis should be based on the comparison between the test procedure's p value and the previously chosen level of the test.
"If the p value is less than the level of the test, then the experimental data are considered to be inconsistent with the null hypothesis, that is rejected, and the result is declared to be significant at that particularly level. If the p value is greater than the level of the test, then the null hypothesis is accepted" [20].

Significance

The term significance literally means the importance of something, and the adjective significant indicates that something is important enough to have an effect or to be noticed.
In statistical inference a result from an experiment in which the null hypothesis is rejected is called significant. In the context of hypothesis testing the concept of statistical significance refers to whether or not the p value of a statistical test exceeds the previously chosen

test level [21]. When a researcher states that a result is statistically significant, it means that it is probably true and not due to chance. Nevertheless a difference or an association may be true without being important: a statistically significant difference is not necessarily biologically or clinically important [19, 21]. On the contrary a result that is not statistically significant could be important.
Consequently it can also be useful to consider the practical significance of a result, with reference to its real importance.

Magnitude of Effect

When a significant result is obtained, not necessarily the observed effect (difference or association) is relevant for practice [22]. Consequently a researcher should establish in advance how big a magnitude should be to consider the effect important.
"The expected magnitude of an effect can be estimated from previously published reports, or if unavailable, may be taken to represent the minimum effect that the investigators would consider meaningful" [19].

Sample Size

A sample is a selected subset of the target population, that is also expected to represent non selected individuals [23]. As a first step, the researcher should establish how many subjects should be studied to detect a significant effect if it actually exists [19].
Sample size determination requires the researcher to quantify the magnitude of the effect that is judged important to be detected, and the allowable magnitude of α and β error [19, 22]. Then the sample size can be calculated by means of simple mathematical calculations chosen on the basis of the statistical test selected for the analysis.
Sometimes in assembling an adequate sample size, objective limits exist such as the low disease rate in the population of interest, logistic problems, or budgetary constraints [19].

Power

The power is the ability of a study to detect an effect (difference or association) if one exists.
The power of a study is influenced by the magnitude of an effect, the sample size, the α and β error. Mathematically, power is $1-\beta$ [19, 22, 24].
The concepts of power correspond to the term sensitivity that is the ability of a test to correctly identify a disease when it is really present; consequently when a test is not sensitive it could indicate sick people as healthy subjects.
Knowing the power of a specific study is a helpful element in interpreting results. This is particularly true for research in which the sample size has objective limitations (for example a rare disease or a small population). In fact if a study is made using an inadequate sample size due to objective limits, non significant results might be due both to a true lack of association and to the impossibility of detecting the association because of the low power. Practical consequences do not take into account this important point, that non significant but relevant results can be easily discarded.
Meta-analysis, aggregating data from single studies, could help to cope with power problems and efficiently furnish a statistical evidence [19].

Confounding & Interaction Effects on P-value

"The term confounding refers to the effect of an extraneous variable that wholly or partially accounts for the apparent effect of the study exposure, or that masks an underlying true association" [16].
With reference to p value this means that when a significant association between two variables is found, this relationship could be modified by the insertion of a third parameter in the analysis: the p value could become smaller or larger, and the association could become even non significant [25].
An interaction between two or more determinants exists at any time that the joint action of these determinants produces an effect that is smaller or greater than the mathematical combination (sum or product) of the individual effects [26]. As far as the interaction's significance is concerned, it is important to highlight that the level of significance should be higher than the usual ones: according to Selvin 0.20 should be the considered level of the test, and a p value less or equal to 0.20 is an indicator for the interaction [26].
In planning a case control study one should regard any known risk factor as a potential confounder and should evaluate any possible interaction between these variables [16].

Confidence Interval

The method of confidence interval, proposed by Neyman and based on frequentist inference, is developed from the assumption that a population parameter is fixed, but unknown [27].
Through data analysis of a sample, it is possible to make inference about the population from which the sample was selected. Clearly the parameter calculated for the sample (i.e. a mean) is not exactly equal to the real value of the population parameter, but it is an estimate of it [28].
A confidence interval for a sample parameter gives an indication of the precision of the estimate showing the range within which the real population parameter is likely contained. Specifically, for a parameter it is possible to construct the lower and upper limits that re-

spectively indicate how small or large the parameter value might be in the population.
The wider the confidence interval is, the more uncertain the parameter estimate is, and on the contrary the narrower the interval the more certain the estimate.
The confidence interval, is calculated with a given probability that the true value of the parameter is contained within the interval. This probability is called the confidence coefficient or confidence level, mathematically it is 1-α, and it is usually considered equal to 95% (1-0.05). In other words there is a 95% chance that the interval you calculate includes the true population parameters [27].
The sample size, the variability (dispersion) among data, and the confidence level influence the lenght of the confidence interval.
A criticism of confidence intervals is that "no distinction is made as to whether the population parameter is likely to have a higher probability of being in the neighborhood of sample parameter compared with being at the ends of the interval" [9].
In accordance with Fisher "the value of the parameter is more likely to be in a neighborhood of sample parameter compared with being located in a neighborhood around the boundary of the confidence interval" [9, 29].
Statistical inference could be made in a more proper way using the confidence interval together with the p value obtained by hypothesis testing. Through the simultaneous use of p value and confidence interval it is possible to assess both statistical and to be oriented about clinical significance.
In fact the p value gives the probability that a sample parameter is obtained by chance: if p value is small the result is statistically significant and probably not due to chance. The confidence interval gives you the probability that the population parameter is big enough to count for clinical or practical significance, giving the smallest and the largest value that the population parameter can assume [29].

p-value according to Fisher

The p value was proposed for the first time by Fisher as part of a method of inference in 1925.
According to Fisher's original purpose the p value was an index useful to measure if observed results are consistent with a null hypothesis: the smaller the p value, the greater the evidence against the null hypothesis [1]. If the p value is between 0.1 and 0.9 there is certainly no reason to reject the null hypothesis; if it is below 0.02 it is strongly indicated that the null hypothesis fails to account for the whole of the facts [8]. Fisher disagreed with the idea that p value was ultimately for the researcher: a p value of about 0.05 should lead to another experiment, not to accept or reject the null hypothesis [1, 8].
Without doubt, in Fisher's approach, the p value has a subjective and gradual interpretation. Clearly in this approach, where the smaller the p value the greater is the evidence against the null hypothesis, it is important to specify the exact p value: a p value of about 0.001 is stronger evidence than a p value of 0.01.

p-value according to Neyman-Pearson

In 1933 Neyman and Pearson proposed a methodology called hypothesis testing, introducing the concepts of alternative hypothesis, type I error, type II error, and power [1].
According to this approach, in research involving a null and an alternative hypothesis, the researcher should establish in advance the magnitude of the effect that is relevant to point out, fix the level of type I error he can accept, and calculate the sample size that at the same time can minimize the type II error and detect the effect if it exists [1, 8]. After the experiment the p value should be calculated through a test of significance; if the p value is smaller than the fixed α level, the null hypothesis should be rejected, otherwise it should be accepted. In Neyman-Pearson's theory not considering α equal to 0.05 but establishing an appropriate α value in advance, is the crucial point.
Thus, in the Neyman Pearson objective approach we decide in advance the α level, and the result of the analysis leads automatically to reject or accept the null hypothesis.
In this metodology, "we make no attempt to interpret the p value to assess the strenght of evidence against the null hypothesis" [8].
For this reason it is important only knowing whether the p value is lower or higher than the α level, and not the exact p value: if α is equal to 0.10 there is no difference between a p value of 0.06 and a p value of 0.01 because both constitute the same evidence against the null hypothesis.
Obviously the researcher is free to change α level, but this must be done in advance of the statistical test [1, 8].

Heirs of Fisher or Heirs of Neyman-Pearson?

The strong disagreement on p value interpretation has its source exactly in the different methods proposed by Fisher and Neyman Pearson. In fact in time the differences between the two original approaches have been neglected, and the result is a considerable confusion that has led to the widespread misunderstanding of the nature of statistical significance [8].
In scientific literature, the current tendency is to reject the null hypothesis when the p value is equal or less than 0.05, otherwise to accept it, without regard to the type the study and to other available evidence [8]. This approach is arbitrary because neither Fisher nor Neyman-Pearson stated that 0.05 was to be the discriminating value to accept or reject the null hypothesis.
In our opinion the selection of the most suitable approach for a specific study (Fisher's subjective or Ney-

man Pearson' objective ones) should be the first step of the statistical inference. The type of study should be a key factor in this selection process; for example Fisher's approach could be useful for analytic study aimed at exploring the effects of the exposure on health, while Neyman-Pearson approach could be useful for clinical trials testing a new drug.
If Fisher's approach is used, p value should only be only a guide to interpret results, and it should be evaluated in the light of several factors. The confidence interval, the sample size, the power, the biological plausibility, the evidence of dose response, and the consistence of the evidence within and across studies should be looked at carefully.
Analogously, when Neyman-Pearson's approach is used, the α level should not be equal to 0.05 by convention, but should be fixed in advance of the statistical test in the light of the same factors previously listed.

References

[1] Goodman S. *P value*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, pp. 3233-3237.

[2] Smith GD. *Uncertainty and significance*. Int J Epidemiol 2003;32:683.

[3] Berkson J. *Tests of significance considered as evidence*. Int J Epidemiol 2003;32:687-91.

[4] Fisher RA. *Note on Dr Berkson's criticism of tests of significance*. Int J Epidemiol 2003;32:692.

[5] Sterne J. *Commentary: Null points – has interpretation of significance tests improved?* Int J Epidemiol 2003;32:693-4.

[6] Stone M. *Commentary: Worthwhile polemic or transatlantic storm-in-a-teacup?* Int J Epidemiol 2003;32:694-8.

[7] Goodman S. Commentary: *The P-value, devalued*. Int J Epidemiol 2003;32:699-702.

[8] Sterne J. *Sifting the evidence what's wrong with significance tests?* BMJ 2001;322:226-31.

[9] Zelen M. *Inference*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, pp. 2034-2046.

[10] Schlesselman J. *Basic Methods of Analysis*. In: Schlesselman J, eds. *Case-Control Studies. Design, Conduct, Analysis.* Oxford: Oxford University Press 1982, p. 174.

[11] Senn S. *A Brief History of Statistics*. In: Senn S, ed. *Statistical Issues in Drug Development*. John Wiley & Sons 1997, p. 19.

[12] Salsburg D. *Hypothesis Testing*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, pp. 1969-1976.

[13] Schork MA. *Null Hypothesis*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, p. 3081.

[14] Clayton D, Hills M. *Null hypothesis and p-values*. In: Clayton D, Hills M, eds. *Statistical Models in Epidemiology*. Oxford University Press 1993, pp. 96-109.

[15] Schork MA. *Alternative Hypothesis*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, pp. 132-133.

[16] Schlesselman J. *Basic Concepts in the Assessment of risk*. In: Schlesselman J, ed. *Case-control studies. Design, conduct, analysis.* Oxford University Press 1982, pp. 27-68.

[17] Armitage P. *Association*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, pp. 202-203.

[18] Hennekens CH, Buring JE. *Statistical Association and Cause Effect Relationship*. In: Hennekens CH, Buring JE, eds. *Epidemiology in Medicine*. Little, Brown and Company 1987, pp. 30-34.

[19] Hennekens CH, Buring JE. *Analysis of epidemiologic studies: evaluating the role of chance*. In: Hennekens CH, Buring JE, eds. *Epidemiology in Medicine*. Little, Brown and Company 1987, pp. 243-271.

[20] Haseman JK. *Level of a Test*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, p. 2231.

[21] Friedman L. *Clinical Significance vs. Statistical Significance*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, pp. 676-678.

[22] Senn S. *Determining the Sample Size*. In: Senn S, ed. *Statistical Issues in Drug Development*. John Wiley & Sons 1997, pp. 169-185.

[23] Last JM. *A Dictionary of Epidemiology*. Oxford University Press 1995, p. 150.

[24] Atkinson AC. *Power*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, pp. 3460-3466.

[25] Breslow NE, Day NE. *General Considerations*. In: Breslow NE, Day NE, eds. *Statistical Methods in Cancer Research.* IARC Scientific Publication 1980, pp. 84-119.

[26] Selvin S. *The analysis of contingency table data: logistic model I.* In: Selvin S, ed. *Statistical Analysis of Epidemiologic Data.* Oxford University Press 1996, pp. 197-242.

[27] Cook NR. *Confidence Intervals and Sets*. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. John Wiley & Sons 1998, p. 861.

[28] Rothman KJ. *Epidemiology, an introduction*. Oxford University Press 2002, pp. 113-129.

[29] Fisher LD, van Belle G. *Statistical Inference: Populations and Samples*. In: Fisher LD, van Bell G, eds. *Biostatistics. A Methodology for the Health Sciences*. John Wiley & Sons 1993, pp. 113-115.

■ Received on June 30, 2006. Accepted on July 18, 2006.

■ This paper is intended mainly for health professionals, who have to deal with significance, in order to identify relevant health determinants, to develop health promotion programmes, risk assessment, to realize risk communication programs, to plan vaccination campaigns, and to evaluate the effectiveness of their actions. Public health and health promotion require sound research methodologies, able to legitimise and justify the action and the financing from a theoretical framework firmly rooted in relevant sciences. Epidemiological research aims not only at explaining natural and social phenomena, but also at promoting and facilitating public health action. Most of public health decisions are hopefully based on epidemiological research, but often the interpretation of data analysis is affected by myths and traditions, not fully rooted in scientific evidence. This aspect is particularly important for the concepts of "p value" and "significance", that are probably the statistical terms most used in epidemiological literature.

■ Correspondence: Prof. Paolo Contu, Department of Public Health, University of Cagliari, via Porcell 4, 09100 Cagliari, Italy - Tel. +39 070 6758362 - Fax +39 070 668661 - E-mail: pcontu@pacs.unica.it