Current perspectives

Frequent mistakes in the statistical inference of biomedical data

Manuel Martínez-Sellés[*§], Luis Prieto[§], Inmaculada Herranz[§§]

[*]Cardiology Department, Hospital Universitario Gregorio Marañón, [§]Biostatistics Unit, Facultad de Medicina, Universidad Complutense, [§§]Statistics Department, Universidad Complutense, Madrid, Spain

Key words:
Data analysis; p value; Statistical inference.

Although there is a plethora of books, reviews and articles defining the meaning of the p value, many investigators make errors when reaching conclusions from their work based on this p value. Most report their data using rigid sentences such as "the results are not statistically significant, p > 0.05", frequently misunderstood by the authors and the readers. Many professionals are not aware of their limitations in this field rendering the problem even more complicated.

In this article we include advice from experts against mistakes frequently made by investigators, such as the plea made by Rothman that a correct interpretation of the data should not be replaced by sentences such as "statistically significant" or "statistically non-significant". Detailed comments on the more frequent mistakes as well as the reasons for their appearance and persistence during the decades are presented. Finally, a comprehensive explanation of the p value is included, to emphasize that to avoid these mistakes there is no need to learn the mathematical basis of tests but that logic alone would suffice.

(Ital Heart J 2005; 6 (2): 90-95)

Received October 5, 2004; revision received December 10, 2004; accepted December 16, 2004.

Address:

Manuel Martínez-Sellés, MD

Cardiology Department
Hospital Universitario
Gregorio Marañón
Dr. Esquerdo, 46
28007 Madrid
Spain
E-mail: mmselles@navegalia.com

Introduction

Experts are continuously reporting investigator's mistakes. Healthcare investigators are constantly using statistical inference: is the result seen in the sample true for all the population? Which results are really valid and which ones are just anecdotes? Until the end of the 19th century there was no support from the probability field to help solve this problem. The 20th century brought a special present to investigators in all the medical sciences. Pearson[1], Student[2] and Fisher[3] developed the bases of statistical inference with two fundamental tools: significance tests and confidence intervals (CI), and, thanks to them, a real assistance in the interpretation of data became available. However, the benefit these resources provided was sensibly reduced by the infrequent use of the CI and the overuse, frequently in an inappropriate way, of significance tests.

In spite of the frequent publication of books, articles and editorials that familiarize investigators lacking a statistical background with the meaning of the p value, its interpretation and limitations, many researchers still make serious mistakes when reaching conclusions based on a p value. Moreover, many investigators have simplistic ideas in this area that are passed on to the next generation with the risk of assuming wrong conclusions not supported by the data. Many healthcare professionals do not understand the logic of statistical inference and are trapped between the lack of knowledge of this method and the imperious need to use it inside the scientific community. This disparity produces sentences such as "the results are not statistically significant as $p > 0.05$", that are often misinterpreted both by the author and the reader. The problem is even more serious as many researchers are not conscious of their gaps in this area.

The frequent misuse of the p value has been lamented by experts since its introduction. As early as 1956 Fisher explained that although a p value of 0.05 (or any other) was useful for decision-making, in commerce and technology, it was inadequate to increase scientific knowledge[4]. In 1971 Armitage[5] extended this thought: "The 5% level has became widely accepted as a convenient yardstick for assessing the significance of the departure from a null hypothesis. This is unfortunate, because there should be no rigid distinction between a departure, that is a p value, which is just be-

yond the 5% significance level and one which just fails to reach it". In 1985 Feinstein[6] added "Farewell statistical significance" ("p < 0.05"), you misleading and treacherous companion. Your adverse side reactions and intracerebral toxicity have become too overwhelming to compensate whatever efficacy you provided ...". In 1998 Rothman[7] was even more explicit: "Proper reasoning requires more than just a classification of each study into significant or not significant". Thus, degradation of information about an effect into a simple dichotomy is counterproductive and may be misleading. Why has such an unsound dichotomous practice become so ingrained in scientific research? Undoubtedly, much of the popularity of these methods stems from the apparent objectivity and definiteness of the pronouncement of significance. Declarations of significance or its absence may supplant the need for a more refined interpretation of data; the declarations may serve as a mechanical substitute for thought. The neatness of an apparent clear-cut result may appear more gratifying to investigators, editors and readers than a finding that cannot be immediately pigeonholed.

Many others among the greatest theoretical and practical biostatisticians have denunciated this erroneous way of proceeding in biomedical research. This flawed interpretation of the significance tests does not reflect incapacity of biomedical investigators but only the poor teaching that most of them received in this field. As statistical analysis is a mathematical discipline, it is usually explained to healthcare professionals using a mathematical language that is unintelligible to most of them. However, and contrary to the widespread opinion, to avoid these mistakes it is not necessary to learn the mathematical basis of the significance tests, as the logic process behind these tests is the same we use in everyday life.

The most frequent and serious mistakes

Research conclusions do not always adequately reflect the results obtained and frequently there are important disparities between the data and their interpretation. Below is a detailed description of the most important mistakes related to a misinterpretation of the p value (Table I).

Not publishing, underestimating or reporting a result, classifying it as "non-significant" due to a p value slightly over 0.05. A p value that is just beyond the 5% significance level and one which just fails to reach it should have a similar interpretation[5], as the truth is that the difference in the mental attitude to be associated with a probability of 0.04 and one of 0.06 should be negligible[8]. Results should never be reported as "non-significant", as this is very poorly informative on the level of reliability of a conclusion. The exact p value must be presented to minimize the risk of misinterpretations. Sentences such "... there was an increase ... the difference failed to reach statistical significance ... there was a gradual decline ..., although the difference also failed to reach statistical significance ... values remained elevated ... p = non-significant"[9] should be avoided.

Table I. The most frequent and serious mistakes in medical research.

Underestimating a result due to a p value slightly over 0.05, reporting it as "non-significant" and not presenting the exact p value.
Reporting the result of a test as "p < 0.05" with no further mention of the p value.
To conclude, when high p values are obtained, "the null hypothesis is true" instead of "the null hypothesis may be true".
Not providing the pertinent confidence interval.
To mistake decision-making, rarely needed in medical research, with the mental process of readjusting the degree of confidence we have in a hypothesis.
To ignore the intrinsic limitations of the tests of significance, forcing clear conclusions (asseverating that an effect exists or does not exist).

To report the result of a test as "p < 0.05" when the p value obtained is, for example 0.04, but also when it is 0.00004. A p value of 0.04, lends modest support to the hypothesis that the effect found in the sample really exists in the general population, whereas a p value of 0.00004 gives almost definitive evidence. It is preferable to avoid the dichotomy – "significant" and non-significant" – by attempting to measure how significant the departure is, that is, by reporting the p value[10], without limiting the information provided to the readers. However, the medical literature is full of sentences such as "the costs increased from 34 ± 3 in a group to 56 ± 3 in the other (p < 0.05)"[11] or "functional class improved significantly from 2.8 ± 0.4 to 1.3 ± 0.4 (p < 0.05)"[12].

To conclude, when high p values are obtained, "the null hypothesis is true" instead of "the null hypothesis may be true". In a classic revision of clinical trials that concluded that there were no "significant" differences between the therapies used, Freiman et al.[13] found that 50 of the 71 trials were consistent with benefits of the new treatment of ≥ 50%. In all of these trials, the original investigators interpreted their data as indicative of no effect because the p value was not "statistically significant". More recent studies show that this problem continues in biomedical literature, with an erroneous interpretation of the results in 50% of studies[14]. Actually, a high p value only means that many hypotheses may be true, including the null hypothesis. Asserting that, precisely the null hypothesis is the one that is true usually is unjustified, as absence of evidence

does not mean evidence of absence[15]. However, this is a very frequent mistake: a recent paper entitled "Hemochromatosis mutations are not linked to dilated cardiomyopathy in Israeli patients" included the sentence "There was a non-significant trend to a difference in the prevalence of the homozygous H63D mutation between the cardiomyopathy patients and the healthy blood donors (3.18 vs 0%, p = 0.076)"[16].

To exaggerate the meaning of the p value, without providing the magnitude of the effect found in the sample with the pertinent confidence interval. This contributes to the confusion between "a very significant" and "a very big" effect. The truth is that whenever possible the magnitude of the effect found in the sample, as well as the CI to that effect in the population, should be reported. An effect may be very significant but clinically irrelevant, as statistical significance frequently does not translate into clinical significance[17].

To confuse decision-making (to take one or another action on the basis of the sample results) with the mental process of readjusting the degree of confidence we have in a hypothesis. The choice between two alternative courses of action, when, as in industrial quality-control activities, a single study forms the sole basis for it may be justifiable. However, medical research pursues not factual decisions: it just incrementally contributes to an existing body of knowledge[18]. Therefore, the purpose of a biomedical experiment is not to precipitate decision-making but to reset the degree of confidence in the validity of a given hypothesis. This confidence should not be a matter of all or none; rather, the scientific work should be focused on establishing reasonable convictions and not on prescribing actions[19].

To ignore the intrinsic limitations of the significance tests, forcing the presence of clear conclusions (declaring that an effect exists or does not exist). The truth is that most studies do not provide a definite answer in favor or against a given hypothesis, and simply classifying the results as "significant" or "non-significant" does not reduce the level of uncertainty. In many cases, publishing medical research does not imply a firm and final decision; it is just a matter of contributing to an existing body of knowledge. Sentences like "dopexamine increases internal mammary artery blood flow", based on a p value of 0.028[20] give the false impression that a previously unresolved question is definitively solved.

In clinical terms we must make the diagnosis that p = 0.05 has become the accepted standard, but it is time for us to say that this is bad practice[21], with adverse side effects and intellectual toxicity too overwhelming to compensate for whatever efficacy it may have provided[6].

Why the value 0.05 is irrelevant

The value 0.05 is not a frontier that points out important differences, and this should be obvious when performing statistical inference as it is in everyday situations including daily clinical practice. The decision to perform or not a surgical intervention would be similar if the mortality related with it were 5.3% (0.053) or 4.8% (0.048). For the same reason, when performing biomedical research, a p value of 0.053 cannot imply a different conclusion than a p value of 0.048.

The p value, as all probabilities, may take any value between 0 and 1, and varies gradually, with no magic number constituting an abyss perfectly delineating two different zones. We have many examples of continuous variables with no separation point in two qualitatively different groups. When registering a patient's age it would be absurd to consider only if he is over or under 40 years. The same applies for weight, blood pressure and other parameters. In all these examples we can imagine situations where it is interesting to classify the patients in two groups; however, this does not mean that the value chosen in that moment has intrinsic properties that make it usable as a definite frontier in the future. We can, and we do, take an arbitrary diastolic blood pressure value (e.g. for example 90 mmHg), as a diagnostic tool with therapeutic consequences. However, it is obvious that a value of 89.9 mmHg is equivalent to one of 90.1 mmHg, but 90.1 mmHg is very different from 120 mmHg.

Reasons for these mistakes

We already know that 0.05 is no frontier and that while very low p values are strong evidence against the null hypothesis, high p values are not evidence in support of the null hypothesis. What are the reasons that have made do that these mistakes persisted for several decades? Some researchers do understand that the value of 0.05 is only a convention with no intrinsic quality. However, they usually add that they have to use some value to decide whether a result is statistically significant or not. This kind of thinking, deeply established in many investigators, implies a conceptual mistake that constitutes the main cause of all the misperceptions in this field. The truth is that, contrary to the general opinion, in most scientific and health settings, there is no need to take an immediate and final decision. Moreover, as Rothman[7] stresses "it is presumptuous if not absurd for an investigator to act as if the result of his or her study will form the sole basis for a decision". Fisher designed the significance tests, at the beginning of the 20th century, precisely to permit the constant resetting of the degree of confidence in the validity of a given hypothesis. This constitutes the basis of scientific investigation.

In some specific circumstances, the p value is explicitly used to make a choice between two possible options. In these cases it is mandatory to select a separation value, so that with a p value below the one selected an action is taken, while with a higher p value another option is taken. For such circumstances, Neyman and Pearson developed the test of hypothesis. A frontier p value, 0.05 or any other, is needed in order to make concrete decisions immediately, according to the result obtained in the sample taken. This is frequently the case in some fields of industry, commerce and technology, but is extremely unusual in biomedical research, as this immediate decision-making has a logical basis which is very different from those of a scientist engaged in a better understanding of reality[4]. It is important to clearly differentiate these two situations and urgent to stop the mistakes that have accumulated over more than seventy years.

There are other causes of the erroneous importance that the p value of 0.05 has been given (Table II):
a) the misinterpretation of some comments by Fisher, such as that in which he stated that is reasonable to think the null hypothesis is not correct when the p value is 0.05. However, Fisher did not give any intrinsic property to this value; his comment could have been just as well applied to, for example 0.06 or 0.03;
b) statistic tables for the different distributions (normal, t-test, χ^2, etc.) usually show 0.05 and 0.01 p values. This is only a convention; values such as 0.04 or 0.005, for example, could also have been used;
c) some biomedical journals include, in their author instructions, sentences such as: "the effect will only be considered significant with a p value < 0.05". This convention could even be reasonable as long as it is clearly understood that it is only a convention, and provided that authors and readers do not conclude that p values < 0.05 mean that the effect found is real in the population and that p values > 0.05 mean that the effect does not exist at all;
d) when someone does not understand what the p value means, it is easy for him to use rigid sentences such as "the result is statistically significant" or "the result is not statistically significant". In so doing an author would be using the mechanical repetition of expressions that he does not understand as substitutes for rigorous thinking[7].

Table II. Reasons for a wrong interpretation of p values.

To mistake the test of hypothesis, designed to make a choice between two possible options, with the significance tests that permit the constant resetting of the degree of confidence in the validity of a given hypothesis.
The misinterpretation of some comments by Fisher.
Statistic tables for the different distributions usually show p values ranging between 0.05 and 0.01.
Some author instructions such as: "the effect will only be considered significant with a p value < 0.05".
When someone does not understand what the p value means, it is easy for him to use sentences such as "the result is statistically significant" or "the result is not statistically significant".

There is no need to learn the mathematical fundamentals of the significance tests to avoid these mistakes

The widespread opinion that complex mathematical knowledge is needed to adequately interpret the p value is not true. To correctly apply the significance tests we only have to use everyday logic. That is why clarifying the significance tests is not only necessary and urgent, but also perfectly possible, as there is nothing "mathematical" or complex in the mental process implied. Mathematicians are responsible for the calculation of the p value, but understanding the meaning of this number is a logical issue, available to all health professionals; for the same reason, we can drive a car with no need to study mechanics and electronics. This logical process can be explained with no mathematical tool, using only common situations.

What the p value means

To explain it we will resort to two simple examples, one with a qualitative response and another with a quantitative one.

Example 1: Is a new treatment effective? Let us assume that we have a disease with a 20% spontaneous recovery and that a new treatment "A" became available. We suspect that with this new treatment the percentage of recovery will be higher than 20%. We use this new treatment in 5 patients.

If the new treatment had no effect, we would expect, theoretically, only one recovery (1/5 = 0.2 = 20%). If we find two recoveries, common sense tells us that this slight increase in the number of recoveries is no clear evidence favoring the new treatment. However, if all the 5 patients recover, what would we conclude? Some could think that, although this datum is suggestive of a favorable effect of the new treatment (the recovery rate is higher than 20%), due to the low number of patients, this result could have been reached by chance, even if the new treatment had no effect. Others would argue that, although it is a small sample, the fact that all the patients recover should be considered as clear evidence in favor of the hypothesis that the new treatment does really increase the percentage of recoveries, because if the treatment was not effective it would be very difficult, only by chance, for all patients to recover.

The p value tells us precisely how probable this extreme result is only due to chance: i.e. if the new treat-

ment is not effective. In this example, the p value is 0.0003, which means that if the new treatment has no effect, the probability of all patients recovering is 3 per 10 000. In this context, the probability is only the proportion in which this would happen, and implies that if we perform this experiment 10 000 times, only in 3 would we find that all patients recover without any treatment. As this is very improbable, the obtained result strongly suggests that the new treatment increases the number of recoveries.

Statistical inference is always performed in this manner: we start with a possible situation (the null hypothesis) and we determine the probability of obtaining data similar to the ones we have in our sample, or even more distant to the null hypothesis than our data. The smaller the p value, the greater the evidence against the null hypothesis.

Example 2: Is the heart of transgenic rats bigger? The mean weight of the mature heart of a certain strain of rats (T0), is 60 g with a standard deviation of 12 g. A transgenic rat strain (T1) is generated and we want to determine whether it has the same mean heart weight. The null hypothesis, H_0, is that the mean heart weight is the same (60 g). Four rats are sacrificed and this sample has a mean heart weight of 84 g. A simple calculation yields a p value of 0.00003, that is to say, if the mean of T1 is really 60 g, out of 100 000 studies like this, only 3 would report a mean heart weight of ≥ 84 g. So, this type of sample is considered strong evidence against the H_0. We sacrifice another 4 rats of a different transgenic type (T2) and we find a mean heart weight of 64 g with a p value of 0.25, which means that if the mean heart weight of T2 is really 60 g, out of 100 studies like this, 25 would report a mean heart weight of ≥ 64 g. This type of sample is considered compatible with the H_0 and provides no evidence against it but neither any evidence in its favor. Finally we sacrifice another 4 rats of a third transgenic type (T3), and we find a mean weight of 72 g and a p value of 0.022, that is to say, if the mean heart weight of T3 is really 60 g, out of 1000 studies like this 22 would report a mean heart weight of ≥ 72 g. Here the evidence against the H_0 is moderate, which inclines us to reject it, but with serious doubts. Let us remember that having found a p value < 0.05 is not definitive proof of anything.

Confidence intervals, a complementary approach

CI are usually more informative and easier to interpret. They indicate the interval inside which we think the population value we want to know lies: in this case, the mean heart weight in each one of the transgenic types. Table III depicts the results obtained in each sample together with the 95% CI and 99% CI.

Table III. Samples obtained from transgenic rats. Non-transgenic rats have a mean heart weight of 60 g and a standard deviation of 12 g.

Type	Sample mean (g)	p	95% CI	99% CI
T1	84	0.00003	74-94	71-97
T2	64	0.25	54-74	51-77
T3	72	0.022	62-82	59-85

CI = confidence interval.

We have 95% confidence that the population mean heart weight for T2 lies between 54 and 74 g and therefore that it may be similar to that of T0, or that it may range from 6 units below it up to 14 units above it. For T3 we have 95% confidence that the population mean lies between 62 and 82 g, that is to say, between 2 and 22 units above that of T0. However, the 99% CI tells us that the mean heart weight of T3 may even be below 60 g, although this is not very probable. The reasonable thing is to conclude that with these data we cannot take a definite position, except to exclude that the mean heart weight of T3 is notably smaller than that of T0. It may be similar to that of T0 or it may surpass it by several units, but not many more than 25.

Number needed to treat, the cost-effectiveness approach

When presenting the data of the clinical research of a new treatment it is always useful to calculate the imaginary number of patients per arm that would result in exactly one less event in the new treatment arm (NTA) than in the placebo/standard treatment arm (PTA). The use of the number needed to treat (NNT) provides the readers with a clear image of the potential clinical and economic effect of the new therapy and is a useful basis to compare the cost-effectiveness of different treatments. In order to calculate it for acute diseases, the reciprocal of the risk difference is used, as $NNT = 1/(risk_{PTA} - risk_{NTA})$. For example, if the 10-day mortality after a myocardial infarction were 7% with a new thrombolytic agent and 10% with the standard treatment (control), NNT = 1/(0.1 – 0.07) = 33: this means that 33 patients have to be treated with the new therapy to save one life. With regard to chronic conditions the reciprocal of the hazard difference should be used: $NNT = 1/(hazard_{PTA} - hazard_{NTA})$. For example, if the hazard of mortality with heart failure were 20/100 patient-years with a new therapy and 30/100 patient-years with the standard treatment (control), NNT = 1/(0.3 – 0.2) = 10; this is the number of patient-years of treatment with the new therapy required to save one life.

Conclusions

Medical researchers often need to use the p value in statistical tests, but few of them do so correctly. Although in most studies, the interpretation of the p value is required to arrive at sensible conclusions, most researchers do not have clear ideas in this respect and cover themselves by using expressions such as "the result is statistically significant as $p < 0.05$", which, often, neither writer or reader understands. The widespread use of what is known as the "5% rule" is particularly unfortunate, as it mistakenly considers this quantity to be the cut-off point between what is valid and what is not. The conclusion obtained after a research is not different for p values of 5.1% and 4.9% and investigators should stop thinking of the 5% significance as having any particular importance. Most researchers refuse to understand the issue fearing that it requires a mathematical understanding which may be beyond them. We have attempted to demonstrate that although the calculation of the p value is a mathematical matter, its interpretation is a matter of common sense comprehensible to all those who wish to understand. It is possible and necessary to create the conditions in which thousands of researchers could stop feeling insecure and feel that even they are able to unambiguously understand and apply this necessary tool correctly.

Acknowledgments

We are indebted to Antonio Martinez, MD, for critically reviewing the manuscript.

References

1. Pearson K. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen in random sampling. Philosophical Magazine 1900; 50: 157-75.
2. Student (Gosset WS). The probable error of a mean. Biometrika 1908; 6: 1-25.
3. Fisher RA. On a distribution yielding the error functions of several well known statistics. Proc Inter Cong Math Toronto 1924; 2: 805-13.
4. Fisher RA. Statistical methods and scientific inference. New York, NY: Hafner Publishing, 1956; 1: 4-80.
5. Armitage P. Statistical methods in medical research. Oxford: Blackwell, 1971: 102.
6. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia, PA: WB Saunders, 1985.
7. Rothman K, Greenland W. Modern epidemiology. 2nd edition. New York, NY: Lippincott-Raven, 1998: 187.
8. Box GE, Hunter WG, Hunter JS. Statistics for experimenters. New York, NY: John Wiley, 1982; 5: 109.
9. Verleden GM, Dupont LJ, Van Raemdonck D, Vanhaecke J. Effect of switching from cyclosporine to tacrolimus on exhaled nitric oxide and pulmonary function in patients with chronic rejection after lung transplantation. J Heart Lung Transplant 2003; 22: 908-13.
10. Armitage P, Berry G. Statistical methods in medical research. Oxford: Blackwell, 1996: 95-6.
11. Leesar MA, Abdul-Baki T, Akkus NI, Sharma A, Kannan T, Bolli R. Use of fractional flow reserve versus stress perfusion scintigraphy after unstable angina: effect on duration of hospitalization, cost, procedural characteristics, and clinical outcome. J Am Coll Cardiol 2003; 41: 1115-21.
12. van der Lee C, Kofflard MJ, van Herwerden LA, Vletter WB, ten Cate FJ. Sustained improvement after combined anterior mitral leaflet extension and myectomy in hypertrophic obstructive cardiomyopathy. Circulation 2003; 108: 2088-92.
13. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med 1978; 299: 690-4.
14. Williams HC, Seed P. Inadequate size of "negative" clinical trials in dermatology. Br J Dermatol 1993; 128: 317-26.
15. Gallagher EJ. No proof of a difference is not equivalent to proof of no difference. J Emerg Med 1994; 12: 525-7.
16. Goland S, Beilinson N, Kaftouri A, Shimoni S, Caspi A, Malnick SD. Hemochromatosis mutations are not linked to dilated cardiomyopathy in Israeli patients. Eur J Heart Fail 2004; 6: 547-50.
17. Khot UN, Nissen SE. Is CURE a cure for acute coronary syndromes? Statistical versus clinical significance. J Am Coll Cardiol 2002; 40: 218-9.
18. Sterne JA, Davey Smith G. Sifting the evidence - what's wrong with significance tests? BMJ 2001; 322: 226-31.
19. Rozeboom WW. The fallacy of the null hypothesis significance test. Psychol Bull 1960; 57: 416-28.
20. Flynn MJ, Winter DC, Breen P, et al. Dopexamine increases internal mammary artery blood flow following coronary artery bypass grafting. Eur J Cardiothorac Surg 2003; 24: 547-51.
21. Moye LA. Statistical reasoning in medicine. The intuitive p-value primer. Boston, MA: Springer, 2000: 68.