European Journal of Internal Medicine 35 (2016) 20–23

Contents lists available at ScienceDirect

European Journal of Internal Medicine

journal homepage: www.elsevier.com/locate/ejim

Review Article

The p-value: A clinician's disease?

F.R. Rosendaal *

Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 28 July 2016
Accepted 8 August 2016
Available online 25 August 2016

Keywords:
Statistics
Tests
Hypothesis
p-Value

1. Introduction

It is remarkable that in the interpretation of scientific evidence, be it as reviewer for a medical journal or as reader, clinicians appear to be the ones most impressed by p-values and statistical significance, or the absence thereof. They attach in their reviews and presentations more importance to it than epidemiologists and statisticians. In fact, the latter groups often actively discourage the use, and certainly the overinterpretation of significance testing, as reflected also in reporting guidelines such as STROBE, which see the presentation of p-values in a scientific article as an option, not an obligation. Since the first authoritative publications forcefully arguing against the use of statistical significance testing appeared in the 1980s and 1990s, it is worthwhile to wonder why p-values are still used, and why especially clinicians are so enamoured with them. In this opinion piece I will explain why relying on statistical testing is philosophically erroneous, how it will often lead to the wrong conclusions, and I will give my layman's psychological explanation why this *p-disease* has a strong predilection for clinicians.

2. History

Formal statistical tests were developed in the early 1900s, in which the most prominent figures were William Gosset and Ronald Fisher. Gosset published nearly all his work under the pseudonym 'Student', since the Guinness brewery in Dublin where he worked did not allow any of his employees to publish at all, and made the exception for Gosset's work only after he had convinced the directors that no other company would profit from his work, and then he still had to publish using an alias. His main purpose was to evaluate agriculture experiments, to be able to compare test fields with varieties of barley or different conditions for their yield. Another use of early statistical tests was for quality control, i.e., when samples from a batch of products, say lamp bulbs, are tested, and a decision needs to be made to test more lamps from that batch or even throw out the lot. It is important to note that, contrary to most research in the biomedical fields, no hypotheses for a particular mechanism underly these tests. It is probably the decisive feature of statistical tests that explains their popularity: a dichotomy of yes and no conforms to medical practice, where physicians have to decide to prescribe a drug or do surgery, or not. Or perhaps its attractiveness lies in the simplicity of a cut-off value that helps decide whether something is true or not. Unfortunately, science is not that simple, and is not about making decisions but about understanding nature.

3. What is a p-value

The p-value is the basis for hypothesis testing, in which there is a null hypothesis and an alternative hypothesis. The null hypothesis typically is that there are no differences, no treatment effects, no risk, whereas the alternative hypothesis is that there are. So, for instance, we are playing a betting game with a coin, and we do a test whether the coin is fair. The null hypothesis is that the coin landing heads or tails is equally probably, with a probability of 0.5. The alternative hypothesis is simply that the coin is not fair, $p(\text{heads}) \neq p(\text{tail})$, or, more simple: $p(\text{heads}) \neq 0.5$.

The experiment would be to do a number of throws, count the numbers of heads and tails, and calculate the probability of that observation given that the coin was fair. Suppose we did four throws and observed 4 times heads: the probability for this sequence would be $(1/2)^4 = 1/16 = 0.067$. When we did five throws, and only observed heads, the probability would be $(1/2)^5 = 1/32 = 0.03$. Intuitively it is obvious that two throws showing two heads are meagre evidence for an unfair coin, while a thousand throws of which each and every one shows heads, is very strong evidence for an unfair coin. The probability of that happening with a fair coin would be $(1/2)^{1000}$ which is a very small probability indeed.

These so-called conditional probabilities (for they are conditional on the null-hypothesis being true) are the p-values used in biomedical research. In other words, the p-value is the probability of observing a certain outcome (five heads in five throws) when the coin is fair.

4. Hume's problem

The Scottish philosopher David Hume (1711–1776) stated that definite proof cannot be attained in empirical open systems, as opposed

* Clinical Epidemiology, C7-P, Leiden University Medical Center, P.O. Box 9600, 2300 RC, Leiden, The Netherlands.
E-mail address: f.r.rosendaal@lumc.nl.

http://dx.doi.org/10.1016/j.ejim.2016.08.015

to deductive closed systems. Mathematics is an example of the latter. The system is closed because all is built on a few premises, and no external information is required to reach definitive proof. To prove Pythagoras thesis on the lengths of the sides of a right triangle: $a^2 + b^2 = c^2$, no experiments need to be done in which thousands of triangles are collected and measured, but a final proof that stands for all times can be written on a piece of paper. In biomedical sciences, as in all empirical sciences, we need data to make inferences about nature, and this will never constitute definitive proof. To use the example of the coin: 1000 consecutive throws of heads would be highly suspicious indeed, but offer no definite proof, for this can happen with a fair coin. In fact, if we threw an infinite series of coins, there would one day be a sequence of 1000 heads. The probability that it happens even when the coin is fair, which is never zero according to Hume, is the p-value.

A bit deeper: p-value and sample size

The p-value is a function of the observed effect and the sample size. When we look for deviations from a 50–50 distribution, say whether men and women are unevenly distributed over a certain profession, we would not need a very large sample to detect a gross unequality. For instance, if the distribution would be 70:30, we would have a high likelihood of picking that up with just a sample of 50 people, and a bit less skewed distribution of 60:40 would require a still quite manageable sample size of 200. However, suppose we would be interested in the tiniest differences, too, such as 51:49; then we would need to include 20,000 people to have a fair chance of seeing that small difference. Again, there are no certainties: no sample size will guarantee that an existing true difference will be detected (cf Hume), but the likelihood to detect it, given a prespecified difference and sample size, can be estimated. This is called the power, related to the type II error (type II error = 1-power). In the examples above this likelihood was set at 80%, meaning that if we did a study with 20,000 people of a certain profession, registered their sex, and in the larger world the difference in men and women in that profession was 51% vs 49%, we would have 80% chance to detect it with a p-value smaller than 0.05; and so 20% to miss it. Again we can buy more certainty by increasing the sample size: if we want to be 90% likely to detect this small difference, we would need to include 26,000 individuals.

This reasoning can also be turned around: with a very small sample size even large effects will not lead to a small p-value, and with very large datasets even the smallest differences will lead to small p-values.

The size of a study is not a scientific truth: it is dependent on resources, time, and the availability of volunteers or patients. This implies that the use of the p-value in statistical tests, it being a function of the sample size, cannot be viewed as scientific.

5. Statistical inference

The statistical inference is the conclusion based on a statistical test, and follows the same reasoning as is used for laboratory tests: we know the distribution of a certain biological parameter in the normal population, i.e., for antithrombin. When a patient has 50% of the normal concentration in the plasma, we know that this is extremely unlikely for an individual from the normal population without a defect allele, and hence we conclude this individual comes not from the normal, but from another population, namely that of people with antithrombin deficiency due to a genetic defect. In the statistical test, the null hypothesis represents the 'normal population', i.e., we assume that there is no association or difference. Then we calculate the probability of the observation or even more extreme observations if there was indeed no association (conditional probability). When this probability is very small, we infer, as in the laboratory test, that our assumption was probably wrong, and that the null hypothesis is not true. In other words, when we throw 1000 heads in 1000 consecutive throws, the probability of this happening is $9.3*10^{-302}$, which is so small that we conclude the coin is crooked. The standard cut-off value to make this inference is 0.05. So, with this cut-off we would also after an experiment of five throws showing five heads ($p = 0.032$) conclude that the coin is false. If we do that, we are wrong 0.032, i.e., 3.2% of the times, which is called the type I error.

6. The logic of $p < 0.05$

There is no logic to it. There is no mathematics or biology that supports a cut-off value of 5%. Probably five was chosen because we use a decimal system, 0–9, and five is half of that. And of course there is no reason to prefer a system with ten digits over one with 12 or four: the only reason we use it is because we have ten fingers.

A bit deeper: several outcomes

In the example above, an extreme is given of only one of two possible outcomes (heads and tails) occurring, but what if both outcomes occur in a series? Suppose we threw the coin a hundred times, and in one experiment observed heads once and tails 99 times, and in the other heads 50 and tails 50 times. It is immediately obvious that in the first we would conclude the coin is unfair, and in the second that it is fair. But what are the probabilities? These can be calculated by simple binomial probability formulas. Observing exactly 1 in 100, has a probability of $100*(0.5)^1x(0.5)^{99} = 7.9 \times 10^{-29}$, which is very small indeed (0.00000000000000000000000000079). Observing 50 of each in 100 throws, has a probability of $\binom{100}{50} * (0.5)^{50} * (0.5)^{50}$. This probability is 0.079. That feels surprising, for we thought if any this outcome would indicate a fair coin, and that if the coin was fair, a 50–50 outcome would have the highest probability. The fact is that this is the highest probability for any possible outcome, for we should realise there are a very large number of outcomes for 100 throws, if we take into account not only the number of heads and tails, but also their order. In total, there are $100^2 = 10{,}000$ different scenarios possible. Taking that into account, a probability of 7.9% for 50–50 is quite high. This is, however, not the p-value in this case, which is defined as the probability of the observation or any more extreme observation, so $P(x \geq 50)$. This is 0.54, so we would conclude there seems nothing wrong.

7. What does a significant p-value say?

A common fallacy is that the p-value says anything about the observation being the result of chance, or even is the probability that the association is not true (i.e., that if $p = 0.05$, there is 5% chance that the observation was due to chance, and 95% probability the association is true). Neither of these is correct: whether chance events occur at all is a question for priests or quantum physicists, but it cannot be quantified, nor can the probability of the truth be estimated. As explained above, the p-value has the anchor of the null hypothesis: we calculate the probability of observing these data if the null hypothesis were true, but we cannot calculate the probability whether the null hypothesis is true or

not. What the p-value really tells us, is how often we will be wrong if we infer there is indeed a difference or association while there is not.

8. What does a non-significant p-value say?

Even less. While a significant p-value gives us some idea about the (un)likelihood of the outcome of the experiment under the null hypothesis, a non-significant p-value has only the interpretation that it is not significant. A common fallacy is that a large p-value would prove that there is no association or no difference. This can simply logically be shown to be untrue, when one realises that the p-value is a function of the sample size. Suppose we did a study on a certain genetic variant and type 2 diabetes, and found that the variant doubled the risk, with a p-value of 0.04. Subsequently, we look separately in men and women, and find that in each group the variant doubles the risk, and now in each group $p > 0.05$. Given the smaller sizes of the subgroups, this is a plausible scenario. However, if one would conclude that $p > 0.05$ proves there is no association, one would find oneself in the paradox that a genetic variant did not increase risk in women, and did not increase risk in men, but did increase risk in people. Although the logic here seems obvious, in some form this fallacy is often seen, for instance with multivariable adjustments: again, a study shows an association between a trait, e.g. BMI, and the occurrence of type 2 diabetes, with a doubling of risk for those who are obese, with $p < 0.05$. Subsequently a multivariate model is used, controlling for a series of variables such as age, sex, ethnicity, alcohol intake and exercise. The result is the same association, still a doubling of risk, but now $p > 0.05$. Those who believe the result has changed, have failed to realise that multivariable analysis is done by stratification, i.e., by making many subgroups, and due to the increased degrees of freedom in the statistical test the p-value will increase, just like in subgroup analyses.

9. Sample size and relevance

The dependence of the p-value on the number of subjects in the study can easily misguide the reader. For instance, two studies into new treatments for a certain disease are published with only the p-values, stating that the effect of the first drug was statistically not significant, while that of the second was. The naive reader might think that the second drug should be used, and the first one dismissed. However, the difference may have been the result of differences in sample size, and in reality the non-significant study may have had a more pronounced and clinically relevant effect, worthwhile to explore further, whereas the large study yielded a statistically significant result, but a clinically irrelevant effect. This is illustrated in Fig. 1.

10. All p-values are different, but some are more different than others

Whereas the p-value itself is dependent on the difference between the groups and the sample size of the study, the interpretation depends also on the prior likelihood of the hypothesis. This again is similar to clinical diagnosis, and intuitively known to all physicians: a small deviation on an ECG of a healthy young boy will impress a cardiologist less than when the ECG is from a 55-year-old overweight man, and the shadow on a lung X-ray of a healthy young girl will have fewer consequences than on the lung photo of a 60-year-old smoker. Why? Because the a priori likelihood of cardiac ischaemia or lung malignancies is very low in healthy children, lower than the probability of a measurement error. This reasoning can be quantified, as it is done in diagnostic algorithms.

For instance, when a diagnostic test with a sensitivity (ability to detect the disease) of 90% and a specificity (ability to detect absence of disease) of 95% is used in a population in which half of the patients has the disease, it will classify 90 out of 100 people with the disease as diseased (true positive), and 95 of the 100 non-diseased correctly as not diseased, and therefore 5 of the 100 without the disease as diseased

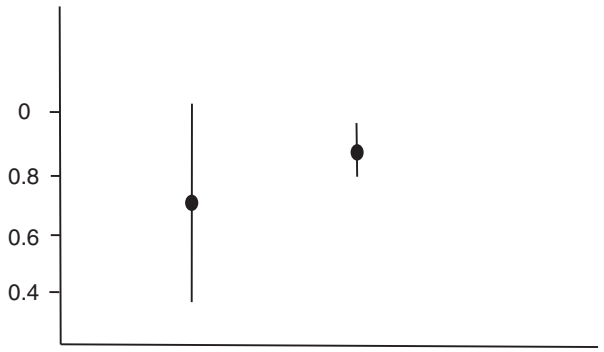

Fig. 1. A statistically significant and a statistically non-significant study. The figure shows two studies with two different drugs against placebo. The study of the left: RR = 0.70, CI95 0.38–1.02, $p > 0.05$. The study on the right: RR = 0.88, CI95 0.80–0.95, $p < 0.05$. The sample size of the study on the right was larger than of the study on the left.

(false positive). Amongst 95 with a positive test, 90 indeed have the disease (94.7%). Here 50% (100:100) is the prior probability and 94.7% the posterior probability. The test may be useful, since most of those with a positive test indeed have the disease. Now we apply exactly the same test in a population in which very few have the disease, say one per 1000. This is the typical situation in screening. Of 100,000 people, 100 will be diseased, and 90 of them have a positive test. Of the other 99,900 without the disease, 5% will be false positive, which is 4995 people. So now, only 90 out of 5085 with a positive test will actually have the disease (1.8% posterior probability), and the far majority of those with a positive test do not have the disease. These two scenarios are shown in Table 1.

This way of reasoning follows the theorem of Bayes, named after an English reverend (Thomas Bayes, 1702–1761), and it can also be used for statistical tests, where we use the prior plausibility or prior belief that a hypothesis is true. The sensitivity to detect disease becomes the power to detect the hypothesis, and the type I error is the false-positive rate (for the type I error was the probability to conclude there is difference (hypothesis is true), when it is not). As Table 1 shows, the reasoning is completely similar to that of a diagnostic test, where the experiment is a diagnostic test of natural truths.

This example shows that the meaning of a statistically significant result is not uniform: like with a diagnostic test it is highly suggestive when the hypothesis that was studied was a priori likely. If, however, the hypothesis was extremely unlikely, even the majority of significant results will not be true. This is the reason why highly statistically significant results of unlikely hypotheses say very little, and why positive trials into the impossible, e.g. a trial into homoeopathy can be dismissed.

11. Ceterum paribus

The basic idea of comparative studies is to study two samples with only one variable differing, and everything else being the same. There is a firm assumption underlying statistical test, i.e., that all variation is random. This is not true in observational studies, and originally statistical tests were seen as appropriate only in randomised trials. But even in randomised trials there are many sources of non-random variation, such as early termination after interim analyses, choices for a particular set of variables in multivariable adjustment, choices for cut-off values, and subgroup analyses. Therefore it is likely that every p-value is an underestimate.

12. Conclusion

Reporting study results by statistical significance can mislead the reader: statistically significant p-values may indicate small irrelevant effects, or may simply be false-positive due to a low prior plausibility. Non-significant p-values cannot lead to any conclusion. The dichotomy

Table 1
Diagnostic and statistical tests for high and low prior probability.

	D+/H1	D−/H0	D+/H1	D−/H0
T+/p < 0.05	90	5	90	4995
T−/p > 0.05	10	95	10	94905
	100	100	100	99,900

The left panel shows a diagnostic test or statistical test when the prior probability of disease or the hypothesis of interest (D+ or H1) is 50%, and the right panel when it is 0.1%. Sensitivity/power is 90% for both, and specificity = 1-type I error is 5%. Posterior probabilities are 90/95 and 90/5085.

in significant and non-significant is unscientific, and is a sign of *hubris*, for a single study can only very rarely be definitive. The cut-off value of 0.05 is fully arbitrary, and there is no reason it should be preferred over cut-offs for significance of 0.10 or 0.01. It is therefore much more useful to present the actual effect size (how much the risk was increased by a risk factor, or how much the endpoint occurrence was reduced by a therapy) with a range of plausible values. This range of plausible values can be represented by a confidence interval, as is shown in Fig. 1.

Conflict of interests

No conflicts of interest.

Suggested reading

[1] Boland PJ. A biographical glimpse of William Sealy Gosset. Am Stat 1984;38:179–83.
[2] Rothman KJ. Significance questing. Ann Intern Med 1986;105:445–7.
[3] Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986;105:429–35.
[4] Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;130:995–1004.