Why the p-value is significant

The p-value can be perceived as an oracle that judges our results. If the p-value is 0.05 or lower, the result is trumpeted as significant, but if it is higher than 0.05, the result is non-significant and tends to be passed over in silence. So what is the p-value really, and why is 0.05 so important?

Are Hugo Pripp
apripp@ous-hf.no

Few statistical estimates have provided so much joy and so much disappointment as the p-value. Its size may have a large impact on such issues as clinical practice, financial conditions for drug manufacturers, and the career and publication success of individual researchers. Since it plays a major role when results are discussed, its widespread use and interpretation remain controversial.

Should we be grateful that the p-value can provide us with a simple answer as to whether our results are significant – or should we abandon the entire concept of significance testing and ban p-values?

What is the p-value?

The p-value is based on two hypotheses. One of these is the null hypothesis, in which it is normally assumed that there is no difference or no effect of a treatment or an exposure. This means that even though the results may show a difference or effect in terms of numbers, we assume that this is caused by random variations and that this difference is therefore not a real, statistical one.

The second and alternative hypothesis is often an assumption that the null hypothesis is untrue. Then a set of mathematical-statistical methods are applied to estimate the probability of what we are observing, given that the null hypothesis is actually correct. The probability of what we are observing or any greater deviations from the null hypothesis – assuming that the null hypothesis is correct – is our p-value. If the p-value is lower than a pre-defined number, the null hypothesis is rejected and we claim that the result is statistically significant and that the alternative hypothesis is true. On the other hand, if the result is not statistically significant, we do not reject the null hypothesis.

So what does statistical hypothesis testing tell us about what we actually want to investigate in a medical or clinical study? In most research studies, the hypothesis is that a treatment or an exposure has an effect. The «medical» null hypothesis is more similar to our statistical alternative hypothesis. Normally, the objective of a medical or clinical research study is to draw conclusions regarding the effect of a treatment or exposure and the magnitude of this effect. Thus, the p-value is not a measure of this probability and provides little information on the medical effect. In other words, the p-value provides information on the probability of our observations, given that the null hypothesis is correct, while the objective of a research study is to provide information on the probability of our medical hypothesis, given our observations.

These two probabilities are different from each other. To complicate matters even further, clinical research nearly always shows a small difference or effect even though it may be neither measurable nor clinically relevant, which means that the null hypothesis is only very rarely true. In reality, hypothesis testing and p-values provide statistical answers to questions that we do not ask ourselves on the basis of assumptions that are not true. This notwithstanding, we use this methodology in most studies – often with major practical benefit.

What the p-value is not

Even though hypothesis testing and p-values are used in medical research, their statistical definition is relatively complex. It might thus be clarifying to discuss what the p-value in fact is *not*. Unless some quite particular assumptions about the data apply, the following is a list of common *misunderstandings* of the p-value (1):

- The p-value is the probability that the null hypothesis is true.
- (1 – the p-value) is the probability that the alternative hypothesis is true.
- A low p-value shows that the results are replicable.
- A low p-value shows that the effect is large or that the result is of major theoretical, clinical or practical importance.
- A non-significant result, leading us not to reject the null hypothesis, is evidence that the null hypothesis is true.
- Non-significant results are a sign that the study has failed.

Why exactly 0.05?

Credit for the choice of a significance level of 5 % is ascribed to the statistician Ronald A. Fisher (1890–1962). Fisher was one of the founders of modern research methodology and statistical analysis. His methods were developed for use in agricultural research and genetics, and have since been applied in a number of disciplines. He is best known for developing analysis of variance and randomised studies (2).

In 1925 he published the book *Statistical methods for research workers*, in which he writes that a significance level of 5 % is an appropriate choice (3): «The value for which P = .05, or 1 in 20, …; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not» (3, p. 45).

We may be left with the impression that a p-value of ≤ 0.05 and the importance of this value in later research may have been caused by Ronald A. Fisher picking a significance level of 5 % more or less at random. If he had chosen 2 %, 7 % or 10 % instead, would medical research and clinical practice have looked different today? Is it true that results and conclusions from large parts of medical research depend on what number a statistician had in mind nearly one hundred years ago?

Although Ronald A. Fisher undoubtedly has had a great impact on the development of trial methods and statistics, it would be simplistic to assign him all the credit (or blame) for this choice of 5 %. Nor is it correct that he chose this level entirely at random; other statisticians were using similar values (4).

Cowles & Davis (5) investigated why Fisher chose 5 % as a significance level. They believe that he was only using what was already an established concept. Karl Pearson (1857–1936), another founder of modern statistics, developed methods for assessing how well data fit with a mathematical probability distribution, which is part of the basis for the frequently used chi-square test of cross-tabulations. He claimed that with a probability of 10 % (i.e. p = 0.1) it is not unlikely that the observed data are random, and further that with a probability of 1 % (i.e. p = 0.01) it is highly unlikely that

Illustration © Espen Friberg/Yokoland

the observed data can be due to random variations. A suitable point between these extremes is 5 %. William Gosset (1876–1937), who developed the t-test, also suggested 5 % as a natural choice of significance level, although he expressed this in other statistical-mathematical terms (4, 5).

Is there anything special about a probability of 5 %? Inspired by their historical investigations of recommended significance levels, Cowles and Davis explored whether there is an intuitive and natural significance level (6). How rarely must an event occur in relation to what is expected before we recognise that the original assumption, i.e. the null hypothesis, is untrue? They provide a simple example. You and your colleague toss a coin to determine who will buy coffee for lunch, but day after day you keep losing. How many days will you be prepared to continue buying coffee for your colleague before starting to suspect that your losses are not coincidental? I would assume that many will be prepared to accept this for four ($p = 0.0625$) or five ($p = 0.03125$) days, but I believe that few would accept that only coincidence is involved if they lose ten days in a row and have to pay for the coffee ($p < 0.001$).

To investigate this systematically, they developed a psychological experiment (6). Volunteers participated in a gambling game. Three cups were place in front of them, and they were told that one of them concealed a small red button. If they chose the right cup, they would win some money. This gamble was repeated until the participants wanted to stop.

For the participants, the intuitive null hypothesis is that they have a probability of one-third for guessing the correct cup in each round of the game. The participants were unaware, however, that none of the cups concealed a red button, and that they thus would lose every time. In other words, the intuitive null hypothesis was untrue. The objective of the experiment was to investigate how many times the participants would repeat the game before starting to suspect that something was wrong, meaning that they would doubt the null hypothesis. More than half of the participants were suspicious after six rounds of repeated losses ($p = 0.088$) and nearly 90 % after eight rounds ($p = 0.039$). The experiment indicates that many people naturally and intuitively will choose a significance level of approximately 5 %.

The future of the p-value in medical research

Some clinical and epidemiological researchers have been highly critical of the use of hypothesis testing and p-values in biological and medical research. Some claim that the entire concept is an outright menace and a threat to scientific progress (7, 8). Other prominent researchers believe that the p-value still fills an important function in statistical analysis of such data, even though they emphasise the importance of confidence intervals and remain sceptical of an absolute rule of 0.05 (9). It is frequently emphasised that reporting of descriptive statistics and effect measurements with confidence intervals are essential and required in addition to the p-value, including in the recommendations for use

of statistics in the Journal of the Norwegian Medical Association (10).

Research is being undertaken with a view to developing methodologies and alternatives to the p-value. Some are working on methodologies based on Bayesian statistics, in which pre-defined assumptions regarding probabilities of effects and results are taken into consideration in statistical analyses. For example, the p-value could be adjusted in light of a medical assessment of the probability of the null hypothesis in a study (11). These assessments are statistically expressed as so-called *a priori* probabilities. These *a priori* probabilities and the actual data from the study are used in the statistical analyses. An obvious dilemma is that of agreeing beforehand what the effect is likely to be and how much each individual p-value should be adjusted.

Genetic statistics face the challenge inherent in having a large number of tests (multiple comparisons). Even though the p-value is used in such analyses, it is not very meaningful to use a 5 % significance level when testing 10 000 genes, for example. If the genes are independent of each other and there is no difference between two groups, one would nevertheless expect 500 significant tests. Special methods have therefore been developed to correct for multiple tests in genetic statistics (12), but systematic correction of all multiple tests in clinical studies remains controversial (13).

The p-value is better is practice than in theory

As noted above, hypothesis testing and p-values provide no direct answers to the questions we pose in most medical and clinical research studies, but this statistical methodology nevertheless remains frequently used. In my opinion, this is because the p-value has proven practically useful. If the probability of our observations is low when we assume that the null hypothesis is true, this is an indirect indication that our observed effects are not caused by random variations. It seems reasonable to undertake an initial assessment of whether this might be the case. I believe that significance testing at the 5 % level is useful as an initial statistical screening before any subsequent assessment of effect size and clinical discussion. Undertaking this assessment has proven useful in practice and is generally accepted, even though in purely methodological terms it has some flaws with regard to its application to clinical data.

The p-value is based on statistical-mathematical methods and as such has little resemblance to a mystical oracle. However, the *Danish Dictionary* describes an oracular reply as «often ambiguous and requiring interpretation by especially competent persons» (14). This description may also well apply to the use of p-values in medical and clinical research.

Are Hugo Pripp (born 1971)
researcher and biostatistician at the Oslo Centre of Biostatistics and Epidemiology, Research Support Services, Oslo University Hospital.
The author has completed the ICMJE form and declares no conflicts of interest.

References
1. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 2000; 5: 241–301.
2. Box JFRA. Fisher, the life of scientist. New York: Wiley, 1978.
3. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925.
4. Curran-Everett D. Explorations in statistics: hypothesis tests and P values. Adv Physiol Educ 2009; 33: 81–6.
5. Cowles M, Davis C. On the origins of the. 05 level of statistical significance. Am Psychol 1982; 37: 553–8.
6. Cowles M, Davis C. Is the. 05 level subjectively reasonable? Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement 1982; 14: 248–52.
7. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010; 25: 225–30.
8. Mitchell MS, Yu MC, Whiteside TL. The tyranny of statistics in medicine: a critique of unthinking adherence to an arbitrary p value. Cancer Immunol Immunother 2010; 59: 1137–40.
9. Vanderweele TJ. Re: The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010; 25: 843–5, author reply 844–5.
10. Aamodt G, Gulbrandsen P, Laake P et al. Presentasjon av statistiske analyser i Tidsskriftet. Tidsskr Nor Lægeforen 2005; 125: 2183–7.
11. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens 2011; 24: 18–23.
12. Montana G. Statistical methods in genetics. Brief Bioinform 2006; 7: 297–308.
13. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998; 316: 1236–8.
14. Den Danske Ordbog. http://ordnet.dk/ddo/ordbog?query=orakelsvar (22.4.2015).

Received 23 April 2015, first revision submitted 28 May 2015, accepted 10 June 2015. Editor: Siri Lunde Strømme.