Data Science in Medical Research

John S Croucher[1], Stephnie Hon[2]

ABSTRACT

Background and Objectives: **A common technique employed in medical research is to answer questions using statistical inference and to calculate an appropriate test statistic and confidence interval that will determine whether some hypothesis should be rejected or otherwise. There are many issues to consider, including the controversial matter of whether a one- or two-sided alternative is appropriate. The objective of this paper is to provide a precise interpretation of a confidence interval as well examining how conditional probability should be interpreted.**

Methods and Materials: **A novel approach is undertaken for explaining the p-value in relation to a confidence interval that has been used to test a medical hypothesis. It is also apparent that even skilled medical researchers have not been sufficiently trained in statistical analysis do not have the ability to interpret statistical results. Examples to illustrate the techniques are provided.**

Results: **The results in this paper will enable the medial researcher to better understand the nature of confidence intervals and to use them more effectively in reaching conclusions.**

Conclusion: **It is known that there are a number of medical research articles appearing in refereed journals that contain erroneous conclusions through the misinterpretation of data analysis. This paper uses a case study to illustrate that their findings can often be misleading or just plain wrong.**

KEY WORDS

confidence interval, hypothesis testing, Bayes Theorem, clinical trials

INTRODUCTION

Invariably, an essential part of being a medical researcher is to perform statistical analyses in various clinical settings. The concepts are vital for an understanding of how the latest research results appearing in learned journals may be judged. It is hard to argue the premise that some medical researchers use, or rather misuse, statistics to justify some hypothesis that they want to 'prove'. The literature is scattered with examples of how misleading interpretations of data can lead to just about any result the user desires (Croucher, 2019). Excellent examples are also provided on this subject (Huff, 1954 and Reichmann, 1964) and it is for this reason that many books highlight how to be aware of these pitfalls.

The experienced medical researcher will have come across, multiple times, situations where a hypothesis has been tested and calculations carried out meticulously, only to stumble at the last hurdle. The case referred to here is when the conclusion is exactly the opposite of what their figures have indicated. That is, it ends up rejecting a hypothesis that is true or not rejecting one which is false. These are known as Type I and Type II errors, respectively.

In this case there is usually no malice intended and the false conclusion has resulted more from a poor understanding of statistics rather than mischievous intent. Indeed, on some occasions a null hypothesis may be rejected, but the null hypothesis itself is stated backwards (really being what should have been the alternative hypothesis) and so again the conclusion is incorrect.

METHODS

Of course, one of the major aims for medical researchers is to obtain a 'significant result'. That is, to discover something unusual or to say that their current theory seems to be correct. and it is here that there is also wide scope for misinterpreting the data, or just plain lying, as some would have it (Johnson, 2006). In some cases, the data itself may have been modified to the extent that observations have either been favourably altered, or non-existent values added to the data to ensure the 'right result'. The ethics of such practice should be a mandatory part of any education in this area.

Other instances include the results of experiments that were never conducted in the first place, typified by the case of a high profile Australian doctor who was struck off the medical register for five years in 1993 for falsifying data on a project (Deighton, 1993). Deficiencies in Australian academic and scientific institutions have been well documented (Martin, 1989) where a number of cases of alleged fraud have come to light. But these are stories that could be told in almost any country where scientific research is undertaken.

A particular type of lying comes with the selection of the nature of the statistical test to be employed. In this regard, one issue is whether to use a one or two-sided alternative. The position is by no means clear cut, with proponents of one-sided alternatives arguing that if one direction of the alternative is essentially impossible or very unlikely then it should be disregarded. For example, suppose a researcher wanted to investigate whether a particular drug, designed to lower blood pressure, really works. That is, it does what it is supposed to do. To use a two-sided alternative would be essentially admitting that there was a chance

Received on November 1, 2020 and accepted on February 2, 2021

1) Department of Actuarial Studies and Business Analytics, Macquarie Business School, Macquarie University
Sydney NSW, Australia
2) Macquarie Business School, Macquarie University
Macquarie Park NSW 2109, Australia
Correspondence to: John S Croucher
(e-mail: john.croucher@mq.edu.au)

that the drug could possibly have precisely the opposite effect, that is, raises blood pressure. If one considered that as a ridiculous proposition and that that the only two possible outcomes are that either it does nothing or lowers blood pressure, then an argument could be mounted for a one-sided test to be used.

There could be persuasive arguments for a one-sided test in such a case, but the opponents strongly disagree, claiming that this is really second-guessing the outcome and that all cards should be on the table and all possibilities allowed for (Moyé and Tita, 2002). Indeed, there are many medical researchers who would always use a two-sided test whatever the circumstances, even if one direction of the alternative was essentially ludicrous.

An interesting question is whether it really makes any difference which one is used. The answer is that often the same result is obtained no matter whether a one or two-sided alternative is employed. But it is those situations where a different, in fact opposite, conclusion is reached that provide the most interest here. And what of those researchers desperately looking for a significant result? To them the choice is clear cut. It is patently easier, using exactly the same data, to reject a null hypothesis using a one-sided alternative than a two-sided alternative, as illustrated in Example 1.

Example 1

Consider the blood pressure case mentioned previously. Suppose a new Drug X designed to lower blood pressure has been developed and a researcher correctly conducts a double-blind clinical trial in which, say, 100 patients are given a placebo and 100 are given Drug *X*, both in the form of a daily tablet. Their blood pressures of both are measured a month after taking their respective tablets.

The mean blood pressure of the Drug *X* group is subtracted from the mean blood pressure of the placebo group. If this difference is large enough then one might conclude that Drug *X* has been effective (in some sense) while, if the difference is small, then Drug *X* may be held to have made no difference to blood pressure.

Suppose that the researcher takes the view that a drug designed to reduce blood pressure will certainly have *no chance* of increasing it. Using this logic, a one-sided test is decided upon. In this case the difference between the two means has to be at least 1.645 standard errors (at $\alpha = 0.05$) for a significant result. But if the conservative view prevails and a two-sided test is undertaken, the difference between the two means would have to be at least 1.96 standard errors for a significant result at $\alpha = 0.05$.

It is clear that, whatever significance level is selected, it is 'easier' to obtain a significant result if a one-sided test is used. And this opens the possibility that an unscrupulous researcher could attempt to justify a one-sided alternative simply with the sole aim that the chances of rejecting the null hypothesis are vastly improved.

The one-sided versus two-sided debate will no doubt continue, but at least the wary reader will understand the possible implications as to why the decision might have been made. There is another troublesome question regarding two-sided test that is often overlooked. This is illustrated in Example 2.

Example 2

Consider the problem in Example 1 regarding Drug *X* that is designed to reduce blood pressure. The researcher, wanting to be open-minded, decides to use a *two-sided* test. The hypotheses would be written something like this:

H_0: The drug has no effect on blood pressure
v
H_A: The drug has some effect on blood pressure

Suppose that a significant result is found to have occurred in that the Drug *X* has caused a mean blood pressure difference of greater than 1.96 standard errors *below* the placebo group. There are two possible ways of framing the conclusion. These are:

(a) As the test was two-sided, the correct conclusion is that the alternative hypothesis must be correct, and that Drug *X* has *some* effect on blood pressure. We can have no opinion on the *direction* of this effect, however.

(b) Even though the test was two-sided, since a significant result was clearly obtained in a particular direction it is obvious that the rejection must have come from this direction. That is, even though a two-sided alternative was used, it makes sense that the conclusion should be in a particular direction. It is therefore reasonable to conclude in this case that the drug did indeed *reduce* blood pressure.

A survey by the authors of medical practitioners as to which of (a) and (b) would be the correct conclusion revealed a surprising difference of opinion that was fairly evenly divided. There are of course arguments for both sides and the matter cannot be settled easily here. But researchers and scientists should at least be aware of the issue and make their own judgements on which avenue they should adopt.

Some of the more common statistical techniques undertaken by medical researchers are those of estimation and confidence intervals, not just for their own sake but for the testing of hypothesis. The purpose of a confidence interval is to provide a range a values for a population parameter such that if a hypothesised value of that parameter were to fall within the interval then there would be no grounds to reject the notion that this value could be correct.

In this regard, consider the case of attempting to estimate a population proportion π and subsequently constructing a confidence interval for it. Of particular interest here is when the event in question is extremely rare, so much so that there is every chance that there would be no successes at all found in a random sample.

One of the basic aims of medical research is to include a section on statistical inference where the problem is to test hypotheses. The principles of Type I and Type II errors are discussed in many textbooks (Croucher, 2016), but these are usually limited to how null and alternative hypotheses may be constructed, along with a recommendation as to which test statistic is appropriate for a particular situation. Most of these books only provide guidelines for conducting a two-sided test without explaining that this is not the only option and may not even be the most suitable. A statistical computer program will provide a confidence interval based on sample data both in numerical form and as a graphical display. However, these have been criticised for not providing the necessary precision and in turn quite useless (Goldstein and Healy, 1995). Other critics (Robinson, 1975) have also expressed unease as to the precise way in which confidence intervals are used in practice, leading to controversy surrounding their appropriate use.

One of the first steps is to select the desired significance level for the test, even before any data are collected. Once an appropriate test statistic has been determined, its value can then be calculated by performing relevant calculations. Depending on the situation, this may either be compared to a critical value or the p-value be ascertained, sometimes aided by a statistical computer program. Depending on the outcome, the null hypothesis is either rejected or not rejected at that significance level.

An alternative and equivalent way of performing such two-sided tests is to construct a confidence interval based on the sample data. In this case, if a test is to be performed at a significance level of α, then a $100\,(1\text{-}\alpha)\%$ confidence interval is formed. The next step is to determine if the hypothesised value of the population parameter lies within the interval. If it does, then the null hypothesis cannot be rejected at α. If, however, the hypothesised value of the parameter does not lie within the interval, the hypothesis is rejected.

An intriguing question asked by researchers is whether it makes any difference as to exactly *where* in the confidence interval the hypothesised value lies. For example, if it lies more towards the centre of the interval can we be more confident of the conclusion than if it lies near the ends? It is routinely taught that the conclusion is clear cut in in that either it lies in the interval or it doesn't, and its precise location is of no interest.

There is, however, much more that can be said apart from such a superficial determination and several thought-provoking questions arise. These include

- If the hypothesised value lies either close to the centre or close to the extremities of the interval, can a stronger conclusion be made?
- If the hypothesised value is way outside the interval, in contrast to being only just outside, can a stronger rejection be made?
- Is it possible to formulate guidelines depending on close the hypothesised mean is to the centre of the confidence interval?

To simplify matters, we restrict the discussion to a set of independent observations taken from a normally distributed population with an unknown value of the mean μ and a known value of the standard deviation σ. The aim is to say something meaningful about the value of μ. Assuming the null hypothesis of μ having a specific value is true, the test statistic has a known distribution and we can determine an interval $(-A(\alpha), +A(\alpha))$ which is symmetric about zero based on the distribution assumed. The interval has probability (1-α) for a pre-specified α lying between 0 and 1. If the test statistic is denoted by S then we know that

$$\Pr(-A(\alpha) < S < A(\alpha)) = 1-\alpha.$$

The next step is to unpick the test statistic and carry out some simple algebraic manipulation to turn the expression:

$$-A(\alpha) < S < A(\alpha)$$

into an interval of the form:

$$\bar{x} - B < \mu 0 < \bar{x} + B$$

From this we note that a confidence interval is simply a random interval. It covers the true mean 100(1-α)% of the time.

Using a z-test for the population mean, the test-statistic

$$S = \sqrt{n}\ \frac{(\bar{x} - \mu_0)}{\sigma}$$

has a standard normal distribution Z with a mean of 0 and a standard deviation of 1. The confidence interval for the unknown value of μ is therefore

$$\bar{x} - z\,(\alpha)\frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + a + z\,(\alpha)\frac{\sigma}{\sqrt{n}}$$

where z(α) is the value of z such that Pr $(Z > z) = \alpha/2$. We will consider three values of α, namely α = 0.01, 0.05 and vv = 0.10, these corresponding to 99%, 95% and 90% confidence intervals, respectively.

RESULTS

It is relatively easy to calculate (or even estimate) the *p*-value of any test statistic and this is usually the recommended approach as it provides greater information than the basic confidence interval approach. With the confidence interval method, it is only possible to state whether the *p*-value is greater or less than a particular threshold. For example, the *p*-value will be less than 0.05 if the test statistic lies *outside* a 95% confidence interval and *greater* than 0.05 if it lies *inside*.

If only a confidence interval is provided, it is impossible to determine the exact *p*-value and so must make our own estimate as to its exact size. Some guidelines for a *p*-value based on the position of the hypothesised mean and its position the confidence interval can be made. That is, for a given confidence interval, the *p*-value can be estimated for any hypothesised mean μ_0.

The idea can be illustrated by constructing plots of the relationship between the *p*-value for a two-sided hypothesis test statistic of a population mean and the position of the hypothesised mean μ_0 in relation to its distance from the centre of the confidence interval. In this regard, *in units of the length of the confidence interval*, plot along the horizontal axis the distance of the hypothesised mean μ_0 from the centre of the confidence interval. For instance, a distance of 0.30 indicates it is 0.30 times the length of the confidence interval away from its centre. In the special case of the distance being 0.50, the hypothesised mean is half the length of the confidence interval away from the centre. That is, the hypothesised mean lies at the extremity of the confidence interval with its corresponding *p*-value being the α-level of the confidence interval. The distances from the centre of a confidence interval for a specific *p*-value will depend on the α-level used.

DISCUSSION

The above result provides another dimension to the notion of using a confidence interval to test a hypothesis. If the calculated sample statistic has a value near to the hypothesised population parameter, it follows that:

(i) the test statistic will have a small value
(ii) The hypothesised population parameter value will lie close to the centre of the confidence interval
(iii) The of the test statistic will have a large *p*-value

In each of these cases the null hypothesis will not be rejected. As the sample statistic has a value that is further from the hypothesised population parameter, then:

(a) The value of the test statistic will become larger
(b) The hypothesised value of the population parameter will tend to be on the outer edges of the confidence interval
(c) The *p*-value of the test statistic will become smaller

If the value of the sample statistic wanders too far, before long it will be outside the confidence interval and then the absolute value of the test statistic will exceed the critical value. In addition, the *p*-value will be less than the selected value of α. This would lead to a rejection of the null hypothesis.

A conclusion based solely on whether a hypothesised value of a test statistic lies either inside or outside a confidence interval is somewhat superficial as much more can be said. Whether it is well inside, just inside, just outside or well outside can provide valuable information. It is unreasonable not to consider this aspect when using such a technique to test a hypothesis as the matter is not simply clear cut.

For example, if a hypothesised value is either just a little outside or a long way outside the confidence interval the null hypothesis is rejected in both cases. It seems reasonable to make a comment of just how strong this rejection really is.

BAYES THEOREM

There is ample evidence of the misunderstanding of statistics by medical practitioners that may well come from a deficiency in their education. An intriguing example is provided by a cognitive psychologist (Gigerenzer, 2003) in which a sample of doctors in the USA and Germany were asked to provide their opinion on the probability that a particular woman, who had a positive result in her mammogram, actually has breast cancer. The information provided to the doctors was:

- The woman is in a low risk group of 40 to 50-year-olds where only 0.8% have breast cancer
- The mammogram will say she has breast cancer, even if she doesn't, 7% of the time (the false positive rate is 7%)
- If the woman has breast cancer, the mammogram will say she has it 90% of the time (the false negative rate is 10%)

One medical practitioner in the survey, who had over 30 years' experience, was presented with these statistics. His conclusion was that there was a 90% chance the woman in question indeed had breast cancer. Others in the sample included 24 German doctors of whom 8 said the chance was 10% or lower, 8 said 90% and the remainder placed the probability somewhere between 50% and 80% chance. Among the 100 US doctors surveyed, almost all, 95 of them, put the chance at about 75%. To determine who, if anyone, is correct, Bayes Theorem can be used as it describes the probability of an event based on the prior knowledge of the conditions that might be related to the event. If the conditional probability is known, the theorem provides the reverse probability.

Suppose there is a hypothesis *B* and the available evidence is *A*. Bayes theorem states that the relationship between the probability of the hypothesis *before* getting the evidence Pr(*B*) and the probability of the hypothesis *after* getting the evidence Pr(*B*|*A*) is:

$$\Pr(B|A) = \frac{\Pr(A|B)\ \mathrm{x}\ \Pr(B)}{\Pr(A)} \qquad (1)$$

where

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|\text{not } B)\Pr(\text{not } B) \qquad (2)$$

In this case, define two events *A* and *B*:

A = the mammogram gives a positive result.
B = the woman really has breast cancer

We are looking for the value of the conditional probability $\Pr(B|A)$. The available data are:

$\Pr(B) = 0.008$
$\Pr(\text{not } B) = 1 - 0.008 = 0.992$
$\Pr(A|\text{not } B) = 0.07$
$\Pr(A|B) = 0.90$
$\Pr(\text{not } A|B)) = 0.10.$

Therefore, using (2):

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|\text{not } B)\Pr(\text{not } B)$$
$$= (0.90 \text{ x } 0.008) + (0.07 \text{ x } 0.992)$$
$$= 0.07664$$

From (1):

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

$$= (0.90 \text{ x } 0.008)/0.07664$$
$$= 0.094 \text{ or about } 9\%$$

The conclusion is that the real chance the woman has breast cancer is 9%, meaning that the vast majority of the surveyed doctors, especially those in the USA, were not even close and utterly exaggerated the probability. About one-third of the German doctors were essentially correct. For women in this age group it leads one to consider just how useful a mammogram really is if the interpretation can be so misleading.

REMARKS

The use of confidence intervals is so prevalent in medical research papers that any addition to the pool of knowledge on their use should be welcomed. This paper has highlighted some of the important and unusual aspects of hypothesis testing that have sometimes been lost along the way in the ordinary explanations of correct procedures. In the instance of one versus two-sided testing, there are certainly those who feel that situations exist where the use of a one-sided test is entirely justified (Peace, 1988).

Another aspect considered is the interpretation of evidence and statistical errors in medical research journals are not uncommon. For example, Olsen (2003) found that 54% of a sample of 141 papers published in the journal *Infection and Immunity* had errors in reporting their analysis. Yim *et al*. (2010) found that 79% of a sample of 139 papers published in the *Korean Journal of Pain* had errors, while Nieuwenhuis *et al* (2011) found that 15% of articles reviewed in the top tier journals *Science*, *Nature*, *Nature Neuroscience*, *Neuron* and *The Journal of Neuroscience* had used the wrong method.

It is only with robust debate on this, and other statistical issues of a medical nature, that the field of medical research can progress and provide stimulating and challenging discussion for those practising their craft. An example of this is the research by Croucher and Hon (2020) that examines the accuracy of clinical testing and how the results may be interpreted.

CONFLICT OF INTEREST

On behalf of all authors, the corresponding author states that there is no conflict of interest.

ETHICAL PERMISSION

None required.

REFERENCES

Croucher, JS (2019). Quantitative analysis for management, 6th edition. McGraw-Hill Australia, Sydney.

Croucher, JS (2016). Introductory mathematics and statistics, 6e (revised), McGraw-Hill Australia, Sydney.

Croucher, JS and Hon, SW (2020). Accuracy of clinical testing and interpretation of results. International Medical Journal, Vol. 27, No. 3, 346-348.

Dayton, L. (1993). Thalidomide hero found guilty of scientific fraud. New Scientist. 27 February.

Fleiss, J.L. (1987). Some thoughts on two-sided tests. Journal of Controlled Clinical Trials, 8, 394.

Goldstein, H. and Healy, MJR (1995) The graphical presentation of a collection of means. Journal of the Royal Statistical Society, Series A, Part 1, 175-177.

Huff, D. (1954). How to lie with statistics, W.W. Norton & Co. Inc.

Johnson, J. (2006). Lying with statistics Part III: the one-sided vs two-sided test.

See URL: <*http://randomjohn.wordpress.com/2006/05/06*>. (Accessed 20 November 2020).

Martin, B. (1989). Fraud and Australian academics. Thought in Action, Vol. 5, No. 2, Fall, 95-102.

Moyé LA and Tita, ATN (2002). Defending the Rationale for the Two-Tailed Test in Clinical Research. Circulation, Vol. 105, No. 25.

Nieuwenhuis, S., Forstmann, BU and Wagenmakers, EJ (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience 14: 1105-1107.

Olsen, CH (2003). Guest commentary: Review of the Use of Statistics in Infection and Immunity. Infection and Immunity 71(12): 6689-6692.

Peace, K. (1988). Some thoughts on one-tailed tests. Biometrics, Vol. 344, No. 3, 911-912.

Reichmann, WJ (1964). The Use and Abuse of Statistics. Penguin, UK.

Robinson, GK (1975). Some counterexamples to the theory of confidence intervals, Biometrika, 62, 155-161.

Yim, KH, Nahm, FS, Han, KA and Park, SY (2010). Analysis of Statistical Methods and Errors in the Articles Published in the Korean Journal of Pain. Korean Journal of Pain 23: 35-41.