Statistics in practice

The simplest statistical test: how to check for a difference between treatments

Stuart J Pocock

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT

Stuart J Pocock
professor of medical statistics

Stuart.Pocock@lshtm.ac.uk

BMJ 2006;332:1256–8

The complexity of statistical methods for analysing clinical data can make interpreting clinical trial reports a daunting task for many readers. However, the key result of many trials could be presented and interpreted using quite basic statistical methods. The overall spirit of this article is to encourage all interested in understanding clinical trials to "feel the data" rather than get too absorbed in the technicalities (and occasional confusions) of advanced statistical techniques.

For many trials the primary outcome is a disease event. This might be death or a composite outcome such as death, myocardial infarction, or stroke. The standard statistical methods—Cox proportional hazard models and log rank tests—take account of variation in patient follow-up times, but the consequent hazard ratios, confidence intervals, and P values seem a mysterious "black box" to some readers. Alternatively, if events relate to a fixed follow-up time then methods for comparing two proportions (for example, the χ^2 test) may be used.

This article describes a much easier method than these, which readers can use to assess quickly the strength of evidence for a treatment difference in an event outcome. It's surprising even how many statisticians don't know this simplest test: I first heard of it from a cardiologist.

The simplest test

Consider a randomised clinical trial with two treatment groups of roughly equal size. Let the outcome of interest be a clinical event.

The key data are the numbers of patients experiencing the event by treatment group. The figure shows how to perform a statistical test of significance based solely on these two numbers.

Calculate the difference in the two numbers of events and divide by the square root of their sum. Call the resulting number z. Under the null hypothesis that the two treatments have identical influence on the risk of an event, z is approximately a standardised normal deviate—that is, it has a normal distribution with mean 0 and variance 1. From commonly available normal distribution tables, z can be converted into a P value (see figure). For instance $z > 1.96$ means $P < 0.05$ and $z > 2.58$ means $P < 0.01$.

This test is approximate but generally gives reliable results for the following reason. With randomisation, the number of patients in the two treatment groups will be almost equal, as will be the length of patient follow-up. Event rates are usually quite low—for example, less than 20% of patients (and often much lower)—and so the number of patients having an event in each group can be considered to have Poisson distribution. Provided that the total number of events is not too small—for example, not less than 20, then the normal approximation for the comparison of two Poisson random variables[1] leads to the formula in the figure.

The simplest statistical test of significance

- Consider a clinical trial with equal randomisation
- Number of events in the two treatments groups are a and b respectively

Calculate $z = \frac{a - b}{\sqrt{a + b}}$

Value of z	P value
1.28	0.2
1.64	0.1
1.96	0.05
2.05	0.04
2.17	0.03
2.32	0.02
2.58	0.01
3.29	0.001
3.89	0.0001

How to perform the simplest statistical test of significance

It takes about 15 seconds to do this test on a calculator, which makes it a useful way of comparing event counts in a trial with equal randomisation. Note that it ignores the numbers randomised and the follow-up times, except to assume they are virtually equal.

The key information lies in the numerators—the numbers with an event—the size of the denominators being unimportant. For instance, if a trial had twice the number of patients (at lower risk) while still having the same numbers of events, the amount of information would be essentially the same. However, doubling the number of events hugely affects a trial's statistical power.

This technique has two limitations. Firstly, if the denominators differ by a non-negligible amount then the test will become biased in the obvious direction. Secondly, if event rates are high the test becomes conservative—that is, P values are larger than they should be. However, for most published trials these potential limitations seem negligible.

This simple test gives an instant feel of the strength of evidence for a treatment difference. However, in any trial publication it should not replace the more

conventional statistical tests, such as a log rank test or (if length of follow-up is fixed) a χ^2 test.

In estimating the magnitude of treatment effect, the relative risk or hazard ratio is usually presented. This is approximately equal to the ratio of the numbers of patients with an event.

Examples using the test

This simple test is now applied to some recent trials (see table). The VIGOR trial reported an excess risk of serious thrombotic cardiovascular events in patients taking rofecoxib compared with those taking naproxen (hazard ratio 2.37, 95% confidence interval 1.39 to 4.00).[2] Unfortunately, the numbers of patients experiencing such events—45 and 19 respectively in the rofexocib and naproxen groups[3]—were not presented in the original publication. On the basis of the formula in the figure, $z = (45 - 19)/\sqrt{45+19} = 3.25$, which gives $P = 0.0012$, which reveals that the treatment difference is very highly significant. This has been subsequently confirmed by a log rank test ($P = 0.0016$). Note the ratio of events $45/19 = 2.37$ is here the same as the hazard ratio.

Data monitoring in the MOXCON trial,[4] moxonidine versus placebo in heart failure, focused on all cause mortality. An interim analysis, with 1860 patients randomised, found 46 and 25 deaths in the moxonidine and placebo groups respectively. The calculation reveals that $z = (46 - 25)/\sqrt{46+25} = 2.49$, thus $P = 0.013$, strong evidence of excess mortality for moxonidine. This key information led to the trial being stopped early. With 73 more patients randomised and 15 more deaths, the final data had 54 versus 32 deaths in moxonidine and placebo groups respectively, log rank $P = 0.012$. At interim analyses, a reliable log rank test is often impossible as for some patients the "last date known alive" is unknown. This simple test is therefore particularly valuable in data monitoring.

This test is also useful in clarifying why different end points in a trial give apparently inconsistent results. In the PROactive trial[5] in 5238 diabetic patients the primary end point (a composite of cardiovascular events) occurred in 514 and 572 patients in the pioglitazone and placebo groups respectively. This difference of 58 events gives $z = 1.76$ and hence $P = 0.078$, similar to the published log rank $P = 0.095$. The authors focused on the main composite secondary end point (death, myocardial infarction, and stroke) affecting 301 and 358 patients in pioglitazone and placebo groups respectively. This difference of 57 events (one fewer than for the primary end point but based on fewer events) has $z = 2.22$ and $P = 0.026$, similar to the published log rank $P = 0.027$. Perhaps the primary end point was "handicapped" into non-significance by including other cardiovascular events that contributed no extra treatment difference. The similarity of findings for the primary and main secondary end points helps to emphasise that $P = 0.095$ and $P = 0.027$ are not far apart in their interpretation. It is sad that achieving $P < 0.05$ carries such unreasonable weight in considering whether a treatment is effective. Both analyses provide modest, but inconclusive, evidence favouring pioglitazone, but the article's post hoc emphasis on the secondary end point merits caution.

Note that the ratios of events ($514/572 = 0.90$ and $301/358 = 0.84$) for the primary and main secondary end point respectively are the same as the hazard ratios using Cox models—that is, simple division achieves the same as more complex modelling.

The simple test is also relevant to meta-analyses, provided that all included trials have equal randomisation. A recent meta-analysis studied the incidence of target lesion revascularisation in six trials comparing paclitaxel-eluting and sirolimus-eluting stents.[6] Crudely combining the data on 3669 patients in all six trials, target lesion revascularisation was done in 95 and 142 patients respectively in the sirolimus and paclitaxel groups. Hence $z = (142 - 95)/\sqrt{142+95} = 3.05$, giving $P = 0.002$. This instant guide to a treatment difference is backed up by the published stratified Mantel-Haenszel test ($P = 0.001$). It is pleasing to see that by ignoring the denominators and pooling the data from all trials, the simple test still gave the "right" answer.

The simplest test can also help interpret claims about treatment breakthroughs in the lay press. For instance, the *Guardian*'s lead article on 10 December 2005[7] had the headline "New cancer drugs put NHS under pressure" and concluded that "pooled results from three European trials of anastrozole (involving over 4000 women) suggested that post-menopausal women who switched from tamoxifen two years after surgery were more likely to be alive two and a half years later. There were 29% fewer deaths among patient who changed." Unusually (and commendably) this article gave the actual numbers of deaths: 66 in those who switched to anastrozole compared with 90 in those continuing to take tamoxifen. Our quick test gives $z = (90 - 66)/\sqrt{90+66} = 1.92$ and hence $P = 0.06$.

Examples of randomised trial results analysed using the simplest statistical test

Trial (treatment comparison)	Endpoint	No of events	Difference	Square root of sum of events	z	Consequent P value	Published P value
VIGOR (rofecoxib *v* naproxen)[2]	Serious thrombotic cardiovascular events	45 *v* 19	26	8	3.25	0.0012	0.0016
MOXCON (moxonidine *v* placebo)[3]	All cause death when trial was stopped	46 *v* 25	21	8.43	2.49	0.013	0.012*
PROactive	Composite primary endpoint†	514 *v* 572	58	32.95	1.76	0.078	0.095
	Death, myocardial infarction, or stroke	301 *v* 358	57	25.67	2.22	0.026	0.027
Meta-analysis (sirolimus-eluting *v* paclitaxel-eluting stent)[5]	Target lesion revascularisation	95 *v* 142	47	15.40	3.05	0.002	0.001

*Published P value in MOXCON based on an additional 15 deaths.
†The composite primary end point included death, myocardial infarction, stroke, acute coronary syndrome, endovascular surgical intervention in coronary or leg arteries, and amputation above the ankle.

Summary points

- Many clinical trials have two treatment groups, equal randomisation, and an event outcome
- The key data are the numbers of patients with the event in each group
- The simplest statistical test compares these two numbers
- It is a useful, quick, and reliable guide to assessing evidence for a treatment difference
- It is one example of how authors and readers need to get a "feel for data" by means of simple, clear statistics

Thus, *Guardian* readers can assess for themselves that this is encouraging but not conclusive evidence of anastrozole's superiority, which helps tone down the article's extravagant assertions.

The value of simplicity

Although many sophisticated, complex statistical methods are appropriately used in analysing medical data, it is important to identify the key information that drives a study's conclusions. In randomised trials with an event outcome, what matters most are the numbers of patients in each treatment group experiencing the event. The simple test described here uses those data alone and gives quick, reliable insight into a trial's core message.

This test assumes that the numbers of events in each group are Poisson variables. Under the null hypothesis, this means that the proportion of the total events in one group is a binomial random variable with probability $\frac{1}{2}$. This and its normal approximation is widely known as McNemar's test.[1 8] However, few texts[1] and indeed few statisticians and medical researchers realise how that test tackles the issue described and is most easily calculated by the formula in the figure.

The test is also useful when inspecting multiple outcomes in a trial—for example, data monitoring committees looking at tables of frequency of serious adverse events by treatment. Rather than calculating P values for every event, they can use the test to identify any interesting numerical differences of potential concern. Although there is an increased risk of "false positives" when studying multiple outcomes, the simple test plus any z score above 2 suggests the difference may not be due to chance.

For clinical trials with equal randomisation and event outcomes this simple test helps put readers in better touch with the key findings. Of course, a more exact test (such as logrank) should be done, but by then the simple test has already given the game away.

I thank Stephen Evans, Diana Elbourne, Tim Clayton, and Joanna Marro for helpful comments on the manuscript.

Contributors: SJP is the sole contributor.

Competing interests: None declared.

1 Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research.* 4th ed. Oxford: Blackwell, 2002:156-8.
2 Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis VIGOR Study Group. *N Engl J Med* 2000;343:1520-8.
3 Food and Drug Administration. *Statistical reviewer briefing document for the advisory committee.* www.fda.gov/ohrms/dockets/ac/01/briefing/3677b2_04_stats.pdf (accessed 2 May 2006).
4 Pocock S, Wilhelmsen L, Dickstein K, Francis G, Wittes J. The data monitoring experience in the MOXCON trial. *Eur Heart J* 2004;25:1974-8.
5 Dormandy JA, Charbonnel B, Ekland DJA, Erdmann E, Massi-Benedetti M, Moules IK, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive study (prospective pioglitazone clinical trial in macrovascular events): a randomised controlled trial. *Lancet* 2005;366:1279-89.
6 Kastrati A, Dibra A, Eberle S, Mehilli J, Suárez de Lezoj J, et al. Sirolimus-eluting stents vs paclitaxel-eluting stents in patients with coronary artery disease. *JAMA* 2005;294:819-25.
7 Meikle J. New cancer drug puts NHS under pressure. *Guardian* 2005 Dec10:1-2.
8 Woolson RF. *Statistical methods for the analysis of biomedical data.* New York: Wiley, 1987:205-13.

(Accepted 21 April 2006)

A memorable patient

Ben

Ben was one of those delightful patients who put up with much but never complained. He had his first heart attack in 1975, but his real claim to fame was in 1985 when he appeared in a cardiology textbook. Seated at a Guy's Hospital lipid clinic, aged 63, he faced the camera to show his xanthelasma and his blood pressure being measured, while continuing to smoke his cigarette.

As time went by, he and I enjoyed that subtle transition, for which general practice still remains the richer, from doctor-patient to professional friendship. Last winter, I was tidying the office before we moved to a new health centre, and I came across the textbook. I managed to persuade Ben to pose for my mobile telephone camera, as he waited for his bus home from the surgery, holding the book open at the appropriate page. With his permission, and the new technology, the pictures were emailed to the author, now a professor of cardiology.

At 84, however, Ben found life getting tougher; he developed heart failure. After emerging from a 10 day inpatient stay minus warfarin and with a better dose of digoxin, he was due to come to me a week later for review. In the event, I was called directly to the car park: after a good breakfast, Ben had got in the car with his wife and son, but shortly into the journey he had complained of central chest pain and lost consciousness. When I saw him it was clear that his intermittent peaceful respiration would soon cease.

"I don't think we are going to do anything are we?" I asked his wife and son. We knew each other, and they agreed calmly just to wait.

Within five minutes, a peaceful end came, but what to do next? I got into his car, and we all drove to the undertaker. We agreed there was much in this which would have appealed to Ben's sense of humour. His son and I laid him out before I returned, only 20 minutes late, for the rest of the surgery.

His death rekindled for me the sense of privilege that comes with continuity of care. I am not sure that this is measurable or worthy of "points" on a so called quality scale, or for how much longer we in general practice may enjoy it.

John S N Anderson *general practitioner, Woodlands Health Centre, Paddock Wood, Tonbridge (dr_nick@talk21.com)*