

# Power tripping: alpha errors, beta errors and power

Geoffrey R. Norman, PhD,<sup>1</sup> and David L. Streiner, PhD, CPsych<sup>2</sup>

<sup>1</sup> Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada;

<sup>2</sup> Department of Psychiatry, University of Toronto, Senior Scientist, Kunin-Lunenfeld Applied Research Unit Baycrest Centre, Toronto, Ontario, Canada

*Frequently, studies don't work out as planned, and they don't find a statistically significant difference—that is, they don't reject the null hypothesis. In these circumstances, the inevitable question is whether there really was no clinically important treatment effect, or whether the study was simply too small (too underpowered) to detect an important difference.*

**I**n our last article (June 2009, p 284), we examined a “study” of an MCAT-boosting “treatment,” the C<sup>4</sup> course. We found that, instead of a mean MCAT of 25.6, the 100 kids in our study had a mean of 27.0. After a bunch of fooling around, we found that there was a likelihood (a *P* value) of 0.015 that this could have arisen by chance if the treatment didn't work. So we rejected the null hypothesis and concluded that the therapy did work.

It need not have turned out this way. For the sake of argument, let's replicate the study with a smaller sample size, say a sample of 49. And after the dust settles, we end up with the same sample mean of 27.0, just by chance. If we go through the same gyrations, we'll find that this is  $(27.0 - 25.6) \div (6.4/\sqrt{49}) = 1.53$  standard errors (SEs) from the population mean; the chance this could happen is 0.126 (two-tailed). So we don't reject the null hypothesis.

What happened? Well, even though the treatment effect was the same—that is, 1.4—the sample size was smaller, so the standard error was larger and the distribution of sample means got wider.

If we didn't have study 1 in hand, an obvious question would be “Did the treatment not work, or did it work, but we just didn't have a big enough sample to detect it?” In short, we're in the awkward position where we have no evidence to decide (yet). While we might be tempted to say it didn't work—ie, to “accept the null hypothesis”—we can't really say that since we did see a treatment effect of 1.4. It's the O.J. Simpson paradox: The jury found him not guilty; they didn't find him innocent. They did not have sufficient evidence to find him guilty, but they did not have enough evidence

to find him innocent either. So he's in a Never-Never Land between guilty and innocent (that is, he *was* in Never-Never Land. They got him the second time around). It's like the verdict used in Scotland: “Not proven.”

## Educated guessing

Back to MCAT (Medical College Admission Test). To get an answer to the question of whether the C<sup>4</sup> course worked, we have to decide how small a difference in MCAT scores resulting from C<sup>4</sup> would satisfy the criterion for effectiveness—that is, how small a difference would we view as still meaningful or useful. Obviously, the smaller the difference we decide is worth detecting, the larger the sample we'll need to get to a *P* value of 0.05. It all depends on an educated guess. A difference of 0.1 is clearly too small to be useful for anything. A difference of 10 points is enough to get anyone into medical school. So our guess at the smallest difference we want to be able to detect is likely somewhere between the two.

Let's suppose we decide that we want the study to identify a difference of two units. Here's the conceptually tricky bit: We want to be sure that if the treatment worked and resulted in an average change of two units, we have a reasonable probability of declaring that it really had a *P* < 0.05.

To work this out, we have to think about all the

Manuscript received May 1, 2009; accepted June 12, 2009.

Correspondence to: Geoffrey Norman, PhD, MDCL 3519, McMaster University, 1200 Main Street West, Hamilton, Ontario, L8N3Z5, Canada; telephone: 905-525-9140, ext. 22119; e-mail: norman@mcmaster.ca.

Commun Oncol 2009;6:322-323, 327 © 2009 Elsevier Inc. All rights reserved.

sample means that would be generated if we were dealing with an effective treatment that, on average, resulted in a gain of two units. This amounts to imposing a second bell curve on the original picture, that we'll call, with great originality, the  $H_1$  distribution, centered on 27.6, with the same standard error as the  $H_0$  distribution,  $6.4/\sqrt{49} = 0.91$ . This is shown in Figure 1.

Now, we know that if the sample mean is too close to the population mean of 25.6, as it was in our  $n = 49$  study, we'll decide that there really was no difference ( $P > 0.05$ ). On the other hand, if the sample mean is far enough to the right, then we'll declare that  $P < 0.05$ . So there is some value of the sample mean where the treatment effect corresponds to exactly 0.05. That's not hard to figure out; we just have to turn the previous equation on its head:

$$\frac{X - 25.6}{6.4/\sqrt{49}} = 1.96$$

since 1.96 SEs corresponds to exactly 0.05 tail probability (two tailed). This is a treatment mean of  $25.6 + 0.91 \times 1.96 = 27.38$ . So any observed treatment mean greater than 27.38 will be declared significant, and any mean less than 27.38 will be doomed to the purgatory of not rejecting  $H_0$ .

Now we can turn it all around. The thing is, that observed sample mean to the left of 27.38 might come from the  $H_0$  distribution. But it could also come from the  $H_1$  distribution. And we can work that out. In short, what's the chance that we will observe a sample mean less than 27.38 if  $H_1$  is true? Just the area of the  $H_1$  distribution to the left of this critical value. So looking at the figure, that's almost 50% of the distribution. The critical value, where we decide that it is, or is not, significant, is at 27.38, and the  $H_1$  distribution is centered on 27.6.

To be precise, the critical value is at  $(27.6 - 27.38)/0.91 = -0.24$  SEs. Look-

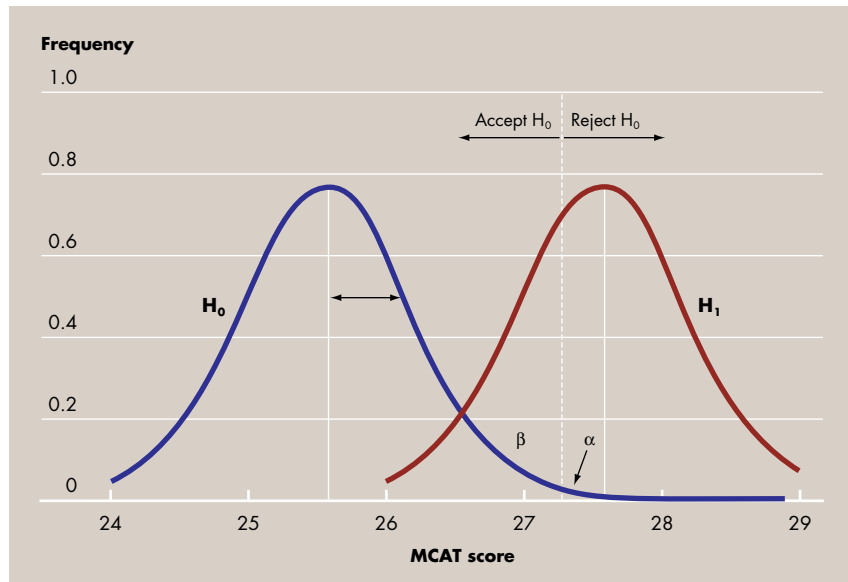


FIGURE 1 Distribution of the null ( $H_0$ ) and alternative ( $H_1$ ) hypotheses

ing this up in the table of the normal distribution, 40.5% of the distribution is to the left. So with a sample size of 49, we have a 40.5% chance of declaring a difference of two units “not significant,” even if it is (if it comes from a distribution where the treatment worked, with a mean of 27.6). This is called  $\beta$  or “beta,” analogous to alpha, the chance of declaring there is a difference when there isn't.

Putting it another way, we can ask what is the chance that we'll conclude there is a difference when there really is. That's just  $(1 - \beta)$ , and is called the power to detect a difference of two units. So it's equal to 59.5%. Just to drive home the effect of sample size, if we stick with the original sample size of 100, power would be 0.88.

**Null or not null**

There is a certain symmetry to the whole exercise. When we do a study we are faced with two options: we either accept or reject the null hypothesis. But we have no way of knowing whether the null hypothesis or the alternative hypothesis is correct. Only the big pathologist in the sky knows that. If we reject the null hypothesis, our treatment mean falls to the right

of the critical value, and we run the risk of making a mistake, a false positive, of alpha. On the other hand, if the mean falls to the left and we fail to reject the null hypothesis, we risk a probability beta of concluding there is no difference when there is one.

It's actually perfectly analogous to diagnostic tests. We could take the same curves shown in Figure 1 and say that, actually, these are the distributions of PSA values for people with and without prostate cancer. Somewhere in the distribution we have established a cutoff or critical value; above that value we'll say a person has cancer and below that, he doesn't. This establishes four zones of pos-

TABLE 1

Relation between the true state (columns) and the decision (rows)

	Truth	
	$H_0$ true	$H_1$ true
Accept $H_0$	True -ve ( $1 - \alpha$ )	False -ve ( $\beta$ ) Type II error
Reject $H_0$	False +ve ( $\alpha$ ) Type I error	True +ve ( $1 - \beta$ ) Power

continued on page 327

sibilities. If a man has a value in the normal range, and we say he doesn't have cancer, that's specificity in the test concept, analogous to accepting the null hypothesis when it is, in fact, true, and has probability  $(1 - \alpha)$ .

On the other hand, if he has a value to the right of the critical value, we'll conclude he has cancer. If it eventually emerges that he doesn't, we have a false positive, or from a statistics perspective, we have falsely rejected the null hypothesis, and the probability is  $\alpha$  that this can happen. Alternatively, he does have cancer, and if his PSA is to the right of the crucial value, we will correctly conclude that he has cancer, which is sensitiv-

ity when referring to tests, and power when referring to statistics, with value  $(1 - \beta)$ . Finally, if his PSA is too low, we will mistakenly conclude he does not have cancer, a false negative with probability  $\beta$ .

The ideas are also formalized in a  $2 \times 2$  table, shown in Table 1.

The table just reiterates what we talked about, but introduces a couple of additional terms: Type I error is rejecting  $H_0$  when it's true (saying there is a treatment effect when there isn't) with probability  $\alpha$ ; type II error is accepting  $H_0$  when it's false (saying there is no treatment when there is) with probability  $\beta$ .

That pretty well summarizes the

underlying logic of statistics. At the end of the day, it's all about estimating the likelihood of particular events—such as a treatment working or not—based on knowledge of the distribution of possible outcomes.

#### ABOUT THE AUTHORS

*Affiliations:* Dr. Norman is Canada Research Chair in Cognitive Dimensions of Clinical Expertise; Assistant Dean, Programme for Educational Research and Development; Professor, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada. Dr. Streiner is Professor, Department of Psychiatry, University of Toronto, Senior Scientist, Kunin-Lunenfeld Applied Research Unit Baycrest Centre, Toronto, Ontario, Canada.

*Conflicts of interest:* None to disclose.