ISSN 1757- 5958

amc technical briefs

Editor: Michael Thompson Analytical Methods Committee AMCTB No 38 March 2009

Significance, importance, and power

It is an important function of statistics to protect us against unwarranted conclusions. This need arises when we are assessing measurement results to see if there has been a change in something, for example the calibration of an instrument or the efficacy of a process. Analytical chemists are all familiar (well, more or less) with the idea of statistical significance. We say that an outcome is statistically significant when it persuades us that the difference of a mean from a reference value is very unlikely to have arisen by chance, given the variability of the measurement results. We are perhaps unduly aware of this significance-testing function of statistics because we make a substantial investment of time in understanding the somewhat intricate rationale behind such tests. Because of that we tend to overlook a simpler but more important aspect of our results, that is, whether any such difference is of an *important* magnitude.

We see this tension between significance and importance rather clearly in nutritional studies. Suppose that we want to test the alleged beneficial effect of eating carrots on a minor disease. We recruit 20 000 volunteers and split them at random into two groups. The control group eat what they normally eat: the experimental group eat in addition a daily portion of carrots. After a period, the incidence of the disease in each group is compared. The outcome is highly significant, so that we know that carrots really work. But the incidence of the disease in the control group is 20% and in the experimental group is 18%. Most people would decide that the improvement was not worth the change in lifestyle, in other words, not important. We note that the criterion of importance comes from an independent external source, not from the data.

Significance

Suppose that we were testing the calibration of an elemental analyser via the reading for nitrogen content obtained when the amino-acid glycine was used as the test material. We can calculate that the nitrogen content of pure glycine is 18.92 %. (This might be relevant, for example, in the determination of protein in foodstuffs.) We repeat the measurement five times and obtain the following results:

18.74, 18.66, 19.01, 18.92, 18.94.

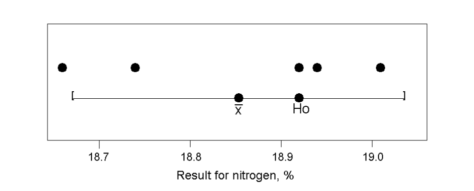
From the statistics (Box 1) we can see that the difference between the mean result and the formula concentration is not significantly different from zero. The *p*-value (the probability, under the null hypothesis, of obtaining a mean result equal to or more extreme than observed) is 0.37, comfortably higher than the usual 0.05 criterion of significance. It seems that there is nothing that requires action. A graphical representation shows this also. (*Note*: it's *always* a good idea to make a graph of the data to check that there is nothing unexpected in the data).

Box 1. Null (H_0) and alternative (H_A) hypotheses and statistics from example data

$$H_0 : \mu = 18.92 \qquad H_A : \mu \neq 18.92$$

$$\bar{x} = 18.854 \qquad s = 0.147 \quad s/\sqrt{n} = 0.0658$$

$$t = 1.00 \qquad p = 0.37$$



Importance

We can't let the matter rest there—we need to know whether the difference is likely to be important, that is, big enough to affect adversely any decisions made on the basis of the result of the machine. Suppose we judge that an error of 0.1 is the maximum tolerable for our purposes. The observed (absolute) difference is |18.85-18.92| = 0.07, slightly smaller than 0.1. At first sight this looks acceptable, but we must be aware of the likely range of possible mean results. The 95% confidence limits tell us that differences between -0.25 and 0.12 would not be rare. So our experiment has been ineffectual—the outcome is not significant, and we don't know whether it is important either.

Suppose that we now decide to do a larger (but not very sensible) experiment with 100 results in order to get a better estimate of the mean value. The mean turns out to be 19.00. The difference estimate is 0.08, but the standard error of the mean ($s/\sqrt{n}$) is now much smaller (because *n* is 20 times greater while *s* will be much the same) and the difference is now significant at 95% confidence. Here we have an outcome that is significant but not important! To ensure an outcome that is useful, we need a minimal experiment but a high probability that a false null hypothesis will be rejected (the power of the test) when there is a difference of important magnitude.

Power

We therefore need to plan our experiment in advance to check whether it would probably give a useful result—a mean result that is significant and important. Assuming that we know the standard deviation of the results, we can calculate the power of the test for a particular number of observations and a given important difference (see Box 2).

> *The power of a test is the probability of rejecting the null hypothesis when it is false*

Some examples of power calculations are shown in Figure 1. Suppose we took the standard deviation of the results as 0.2 % (mass fraction) and, as before, regarded a difference of 0.1 as important. Then we would need 33 repeated results for a power of 0.8, *i.e.*, to see a significant difference in four experiments out of five. If the precision were better ($\sigma = 0.1$), we could get the same level of power with 8 results. Of course, in an example such as that given, we may have little or no control over the precision. It might be necessary to reconsider our ideas about the magnitude of an error regarded as important.

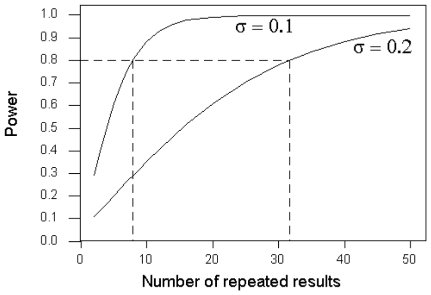


Figure 1. Power of a one-sample two-tailed *t*-test for a difference of 0.1. (See Box 2 for the explanation of how power is calculated.)

This Technical Brief was prepared for the Analytical Methods Committee by the Statistical Subcommittee under the chairmanship of M. Thompson.

CPD Certification I certify that I have studied this document as a contribution to Continuing Professional Development.

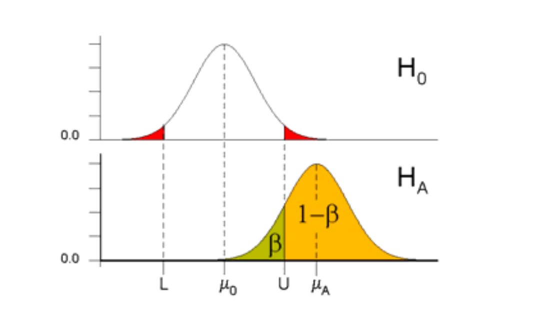
Name……………………..
Signature……………………….…….Date…………

Name of supervisor…………………………………..
Signature……………………….…….Date…………

Other AMC products can be found on:
www.rsc.org/amc/

Box 2. Power calculations

Consider a one-sample two-tailed *t*-test of a mean at 95% confidence, where we have recorded *n* results by a method with standard deviation σ . Under the null hypothesis $H_0 : \mu = \mu_0$, we consider whether the mean $\bar{x}$ of our results would be likely to occur if they were a random selection from a normal distribution with mean μ_0 and standard deviation $\sigma/\sqrt{n}$. The hypothesis is rejected (*i.e.*, we find that there is a significant difference) when either $\bar{x} < \mu_0 - 1.96\sigma/\sqrt{n}$ or $\bar{x} > \mu_0 + 1.96\sigma/\sqrt{n}$. These zones of rejection are shaded red below, with the limits labelled *L* (lower) and *U* (upper) respectively.



We can calculate the power of the test only under *particular* alternative hypotheses such as $H_A : \mu = \mu_A$. Then, if our value of $\bar{x}$ were less than *U*, the null hypothesis H_0 would be accepted with a probability of β . Thus β is the probability of false acceptance of H_0 , and can be evaluated from a table of the normal distribution as the area under the H_A curve to the left of *U* (shaded green). The power of the test under H_A is simply $1-\beta$, the area shaded orange in the Figure.

(Note that if μ_0 were close to μ_A , but still greater, probabilities below *L* would have to be considered in addition. An argument that is a mirror image of the above applies when $\mu_A < \mu_0$.)

The power of a test for a given confidence level depends on σ , *n*, and $\Delta = |\mu_0 - \mu_A|$, and any one of these variables can be calculated from the others. Most often we need to know the power as a function of *n*, given a method with standard deviation σ and a deviation Δ regarded as of important magnitude. This enables us to design an experiment that will provide a useful result on nearly every occasion.

Power calculations are available in statistical software. They can be applied to all kinds of statistical test, including two-sample tests and analysis of variance, and can be used to compare the performance of two or more test methods that have the same aim.