

## EDITORIAL

# Statistics, probability, significance, likelihood: words mean what we define them to mean

Gordon B Drummond<sup>1</sup> and Brian DM Tom<sup>2</sup>

<sup>1</sup>Department of Anaesthesia and Pain Medicine, University of Edinburgh, Edinburgh, UK, and

<sup>2</sup>MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Cambridge, UK

### Correspondence

Gordon B Drummond, Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, 51 Little France Crescent, Edinburgh EH16 4HA, UK. E-mail: g.b.drummond@ed.ac.uk

This article is being simultaneously published in 2011 in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*.

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Brian Tom is in the MRC Biostatistics Unit of the Institute of Public Health in Cambridge, and an affiliated lecturer in the University of Cambridge Statistical Laboratory.

This article is the fourth in a series of articles on best practice in statistical reporting.

### Key points

- 'Student' was a statistician who worked in quality control.
- The *t*-test asks 'how probable are these samples, if they have been sampled from the same source?'
- Considering what would be found when repeated samples are taken is the *frequentist* approach to statistical testing.
- Student's *t*-test does not indicate the probability that the null hypothesis is true.
- Other methods of testing experimental results can be more appropriate.
- We may need to ask 'how different?' or 'is there no effect?'

Statisticians use words deliberately and specifically, but not necessarily in the way they are used colloquially. For example, in general parlance 'statistics' can mean numerical information, usually data. In contrast, one large statistics textbook defines the term 'statistic' to denote 'a characteristic of a *sample*, such as the average score', and the word 'parameter' to denote 'a characteristic of a *population*'. However, for statisticians, statistics means more than just numerical information or a characteristic of a sample. Statistics is also a discipline. It is relevant to all areas of scientific enquiry, and spans study design, data collection, developing methods of analysis and analysing data, interpreting results, and making predictions. Other words that statisticians use carefully, and very particularly, form the cornerstones of statistical reasoning, words such as probability, significance, likelihood. Why does this matter? Well, it may well explain why non-statisticians struggle with the ideas and concepts

used by statisticians. The reader is confused by the terminology, not the least because the statistician uses familiar words in different ways, a jargon similar to other professional vocabularies.

This necessary precision of language becomes evident when considering one of the statistical methods – the *t*-test – commonly used by biological scientists today. It is worth noting to begin with that Student's *t*-test did not originate in the laboratory, but in the factory. The eponymous Student used his pseudonym to conceal his real name because he worked for a brewery, and his publications had to remain anonymous. He used statistics to measure the quality of the product. In one respect, the things that biologists use in some experiments, such as cell cultures and enzymes, are quite similar to a vat of beer. However, there are also fundamental differences. One of these differences is emphasized later in a quote from Student's famous paper. Another is that the brewer wants to be sure of a uniform product. In contrast, the biologist often wants to show that an experiment has generated a change: a positive result!

Student's interest lay in assessing the characteristics of a biological material (such as the raw materials used for the beer). He developed theories to allow him to use measurements of a small sample to estimate the characteristics of the larger amount, for example when he took a small sample from a sack of barley. The overall feature of the sack, such as the average starch content of the grains of barley from the sack, would be termed a 'population parameter'. His groundbreaking paper was 'The probable error of a mean' (Student, 1908). In it, he writes:

it is sometimes necessary to judge of the certainty of the results from a very small sample, which itself affords the only indication of the variability. *Some* chemical, *many* biological, and *most* agricultural and large scale experiments belong to this class, which has hitherto been almost outside the range of statistical enquiry (our emphasis).

Student's contribution was to determine how best to make inferences about a large population on the basis of statistics computed from a small sample. He developed a test of statistical *significance*. As an aside, his use of *probable error*, which was a common expression in the early 20th century, did not refer to the *probability* of making an error: it was used to indicate the variability of his statistic, showing again how we should use words carefully.

It's unlikely, but let's suppose he measured the starch content of each grain in a handful of barley, taken from a sack delivered to the brewery. He wishes to know how well the mean value he obtained from his handful of grains (the sample) will reflect the mean starch content of all the grains in the entire sack. He was aware that in a small sample, randomness would introduce some variation, and also that imprecision in the measurements would also play a part. His method of judging how 'certainly' a small sample can provide a measure of an entire population was later developed in the classical hypothesis testing theory of statistics.

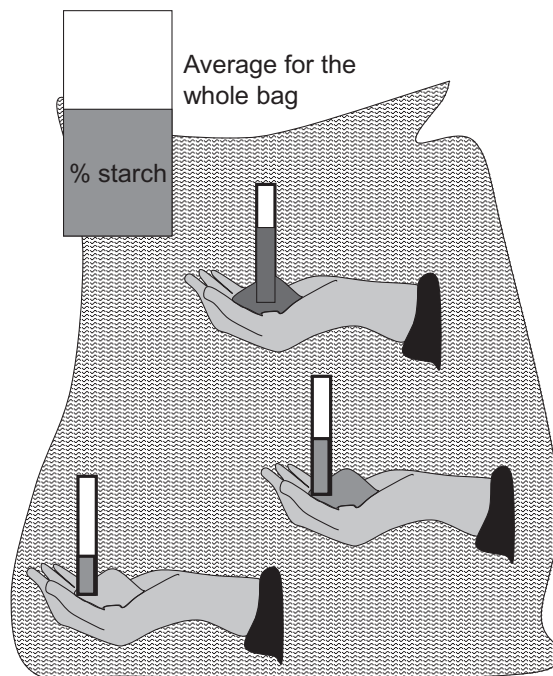
He actually didn't know the mean starch content of the entire bag of barley, one of the *parameters* of the population. To know that would require the measurement of each barley grain in the sack. However, the theoretical starting point of this approach is that we have a population with features that are exactly known, or that these could be determined if necessary. Let's suppose that Student knew that the grains in a sack of barley would have a particular mean starch content. He could then compare this already known value with what he might possibly obtain if he took a small sample of grains at *random* from the sack. He considered the difference between the known mean starch content of the sack and the mean starch content of barley grains obtained in the sample. He could then calculate the degree of certainty of obtaining such a difference. If he took another, different, sample he could calculate another value for this probability. Each probability is associated with a difference between the real mean and the mean obtained from each random sample (handful). Student's contribution was to characterize the probability distribution of these sample averages by taking account of the uncertainty in the sample standard deviations. Most of the time we would expect the mean computed from a sample to be fairly close to the mean for the entire bag of barley (Figure 1).

The concept of taking repeated samples, each yielding a difference between the estimated value from the sample and the true population value, is the basis of the *frequentist* approach to statistical reasoning.

In Figure 2, we show a population of 5000 values, plotted as a distribution histogram. The mean of the population is very close to 1000. Around it are ranged some random samples, each of 250 values, drawn at random from this population. The means of these samples vary from sample to

## Student's distribution

If we take random samples from a bag with these known (or defined) characteristics



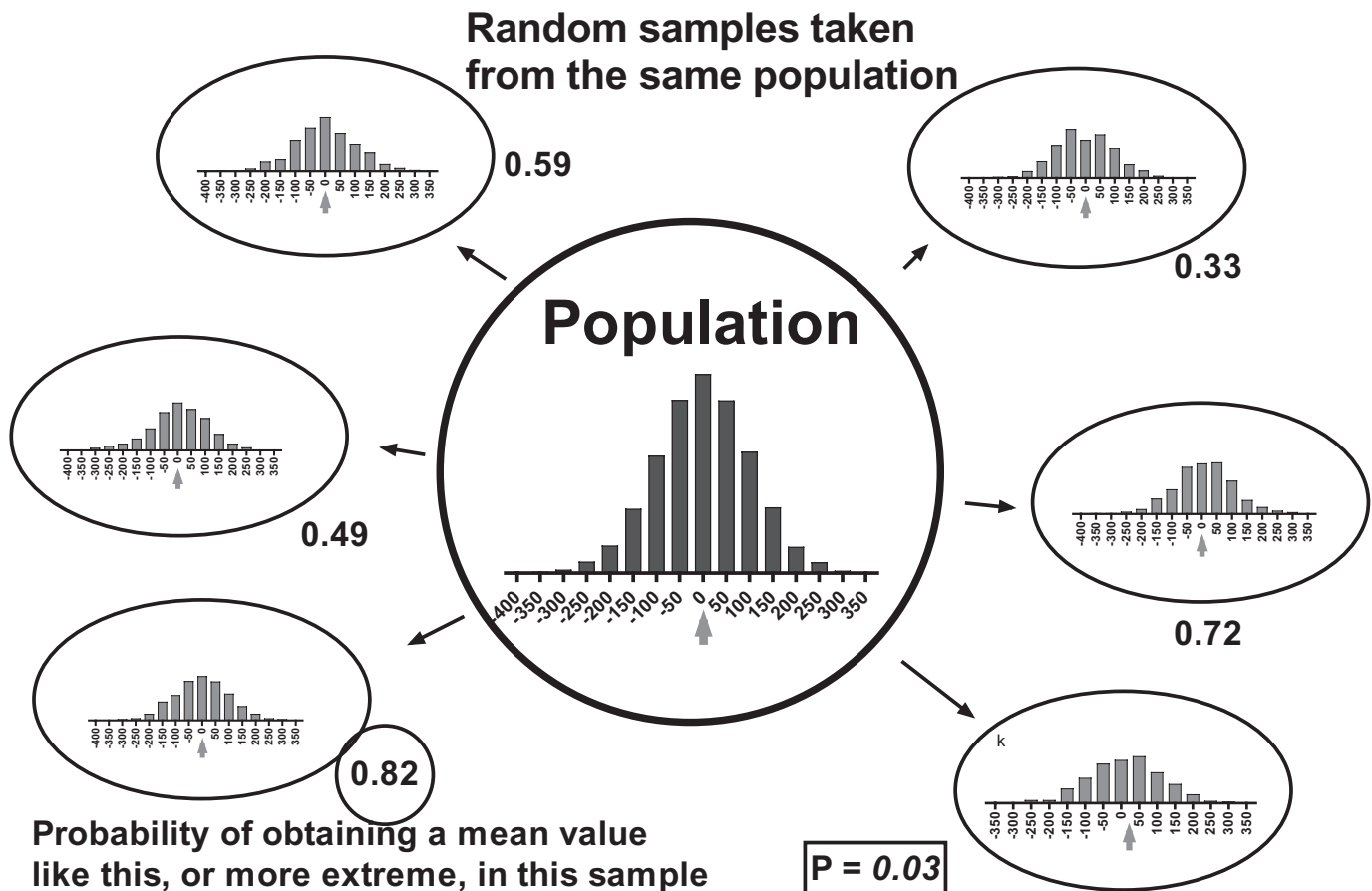
We can calculate the probability of obtaining estimates with averages like these

### Figure 1

A population is randomly sampled and the sample is measured. Student's contribution was to characterize the probability distribution of obtaining samples with any particular estimate. This distribution took account of the uncertainty in the sample standard deviations. Frequently samples would have their mean close to the population mean. One would not often find a sample with a mean that was very different from the actual mean of the whole population.

sample. The probability of finding each sample mean value (or a value further from the population average) in each of these samples can be estimated using Student's procedure. The possibilities are shown and range from 0.82 to 0.03, in other words from close to 80%, down to less than 3%.

When Student published his paper 'The probable error of a mean', he showed how to characterize the possibility of obtaining particular sample means, when repeated samples had been taken from a population with a known mean. More realistically, what he would probably be doing in the brewery would be taking samples from successive bags of what should be a uniform supply, because he wanted a uniform quality of product. If he then found a sample with a mean value that was so extreme that it was improbable, he'd be worried: has the quality of the barley changed? If the samples have mean values that are unlikely, what does this imply? The 'classical hypothesis testing model' was developed to test this concern: is the population mean the same as the standard expected?



**Figure 2**

A population of barley grains summarized as a frequency distribution histogram. The values for each grain are related to the average starch content, showing that some grains have more and some less than the average starch content. Placed around the population distribution we show the distributions obtained from random samples that have been taken from the population. Notice that these samples vary more, because the effects of random variation are more obvious when there are only small numbers in the sample.

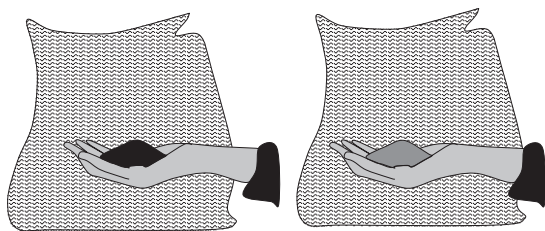
A later development of Student's *t*-test by R. A. Fisher then allowed comparison of the means of two populations, using a similar line of reasoning, shown in Figure 3. We have drawn random samples from each of two sacks. Using the Student's *t*-test, we can judge more formally whether or not the two samples have come from a single population. This can be done by using the measurements of the samples themselves. We are able to calculate, for example, the probability that a particular difference (or something even more extreme) between the sample mean values would have been found in this experiment, if the populations from which they had been drawn had the same mean.

If we assume that the means of the two populations are equal, and then obtain only a small probability that these samples could be possible, then we must conclude one of two things. We must conclude that either the samples do not support the null hypothesis we have made (that they were drawn from populations with the same mean), or we conclude that our sample has been a rare event. As we have pointed out above, it's more than likely that most experiments do not start with the aim of showing 'no effect', although the readers of a research report may be entitled to

ask that the effects that are claimed should be proven. Student's *t*-test, although valuable, doesn't necessarily tell us what we may really want to know. For example, it does NOT give the probability that the samples have in fact come from the same population. Each sample we take will probably have an average starch content that is different from the next. So, Student's *t*-test doesn't give the probability that the null hypothesis IS true. Neither do we have definite evidence that it ISN'T true, nor do we hold evidence that some other alternative possibility IS true. What we do have, using this frequentist system of deduction, is a probability of how consistent our sample estimate of the mean is with an *a priori* chosen value for the population mean (such as the value chosen in our null hypothesis). We can say how frequently we would have seen as extreme a value for our test statistic, just by chance, if we were able to draw repeated samples, AND the *a priori* value chosen (null hypothesis) was in fact the true condition.

The Student's *t*-test merely quantifies the 'Lack of support' for no effect. It is left to the user of the test to decide how convincing this lack might be. A further difficulty is evident in the repeated samples we show in

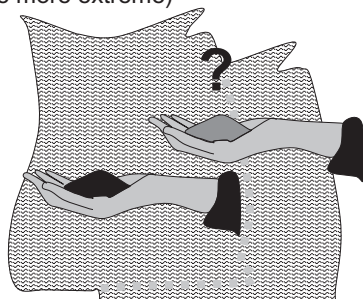
To see if these two samples could be different



We suggest that it's possible that these samples could have come from the same bag

## The *t*-test question is stated:

What is the probability of finding samples such as these (or an average more extreme)



**If** they had both been randomly taken from here?

### Figure 3

The reasoning behind the Student's *t*-test. We have two samples, randomly taken from two populations (different bags). To assess the possibility that the mean in each bag could be different, we hypothesize that these samples are in fact random samples from the SAME population (or more precisely that they are samples from populations with the same mean, which we have taken as having the same, but unknown, standard deviation): this is the null hypothesis.

Figure 2: one of those samples was quite improbable because the *P*-value was 0.03, which suggests a substantial lack of support, but that's chance for you! A parody of this effect of multiple sampling, taken to extremes, can be found at <http://neuroskeptic.blogspot.com/2009/09/fmri-gets-slap-in-face-with-dead-fish.html>

When fully applied to an experiment, the 'classical hypothesis testing model' (developed later on by Neyman and Pearson) involves up to five successive steps of reasoning, and involves a double negative. Despite this, this process remains the primary means by which biologists interpret their data, and it is the approach taken by most textbooks, and also remains the procedure expected by many journal assessors. One large textbook illustrates the several and substantial errors that even trained scientists can make when asked to interpret a result of a Student's *t*-test (Sheskin, 2007). Nevertheless, the book then goes on to state 'when used intelligently the model is extremely useful, and to date has been extremely productive in generating scientific knowl-

edge'. The dominance of these frequentist methods is quite likely to be historical chance, and there is no doubt that these procedures are those expected to be used in most biological science papers. Nevertheless, it can be unnerving to state that the first aim of the process is to describe the probability that one's data, or more extreme values, could have been obtained if in fact there had been no effect in the experiment. Most of the time this is not what most scientists want to know. The question is more commonly 'What effect is there?' or less commonly 'Can I be sure there has been no effect?' and neither of these questions are substantially addressed by the *t*-test.

There are other ways to look at data that will be addressed later in this series. Some provide useful alternatives for scientists. For example, non-parametric tests could be considered. Another possibility may be permutation tests (Ludbrook and Dudley, 1998). (For large samples, this type of test needs a lot of calculation, and this method only became practical with the advent of computers.) A natural extension to significance testing is estimation and, in particular, the use of confidence intervals, where a correspondence with significance testing can be made.

Laboratory scientists do not often take random samples of large populations (although they often assume that their samples represent the population). They conduct experiments on a small amount of material, and get results. These results are used as their starting points, the givens, and are used to inform decisions about populations (assuming that the lab sample is a valid reflection of the population as a whole). Frequently, this requires some prior belief or knowledge about the conditions or features of a population, which can then be modified by the experimental results. This approach, based on Bayes' theorem, is attractive because it tackles the question of the probability of an effect, or no effect, or the probability of the hypothesis given the data. Using a Bayesian framework, further concepts are needed. These include *likelihood*, which is the probability of the data as a function of parameter values, but considers data fixed and the potential parameter values varying. Another important factor is the *prior* condition, and there is no consideration of repeat sampling. In conclusion, it is worth noting that Student's interpretation of his results in his 1908 paper was in a Bayesian, and not frequentist, spirit, although the approach used was one of pure significance testing.

## Conflict of interest

The authors state no conflicts of interest.

## References

- Ludbrook J, Dudley H (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *Am Stat* 52: 127–132.
- Sheskin DJ (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC: Boca Raton.
- Student (1908). The probable error of a mean. *Biometrika* 6: 1–25.