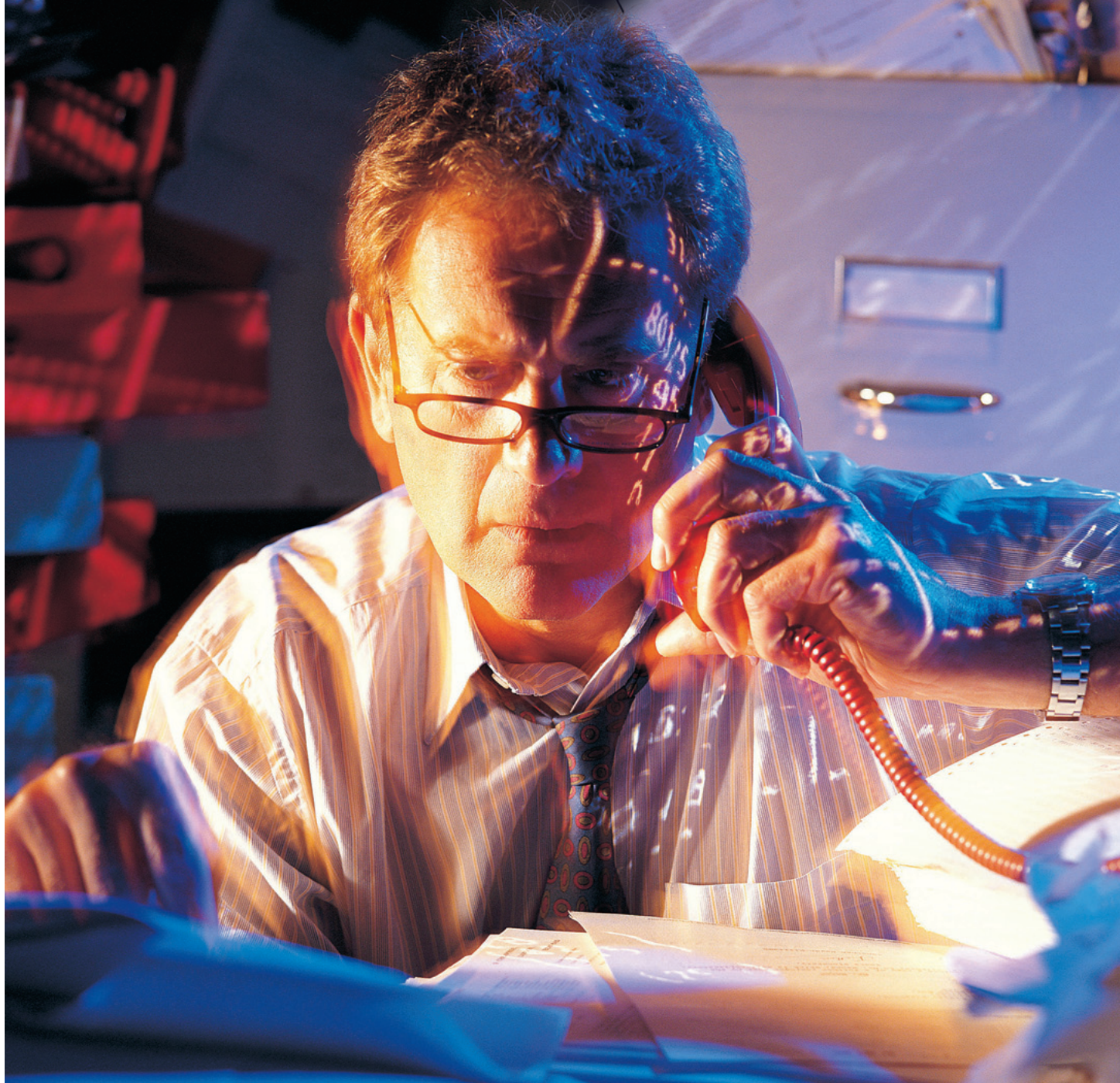


Testing the difference between two groups



In the third of a series of articles about statistics for biologists, **Anthony Hilton** and **Richard Armstrong** test the difference between two groups

Stat Note 3

IN STATNOTE 1

(*Microbiologist*, June 2005), we

described a test to determine whether a sample of measurements came from a normal distribution.

If a variable is normally distributed, it is referred to as a 'parametric' variable. In Statnote 2 (*Microbiologist*, September 2005) we showed that if a variable is parametric, then a sample of observations can be described by the sample mean 'X*' ('central tendency') and standard deviation (SD) 's' ('spread').

We also described two statistical procedures based on the normal distribution.

Firstly, whether an individual measurement is typical or atypical of a larger population of measurements and secondly, we showed that the mean of a small sample of measurements also comes from a normal distribution, viz. that of the population of sample means. The degree of spread of this distribution can be described by the standard error (SE) of the mean. This information was used to calculate a confidence interval (CI) for a sample mean, which is a measure of the error associated with a sample mean as an estimate of the 'true' population mean. In this Statnote, these statistical ideas are extended to the problem of testing whether there is a statistically significant difference between two samples of measurements.

The scenario

We return to the hypothetical experiment described in Statnote 2 to investigate the efficacy of a novel media supplement in promoting the development of cell biomass. To recapitulate, two 10-litre fermentation vessels were filled with identical growth media with the exception that the media in one of the vessels was supplemented with 10ml of a

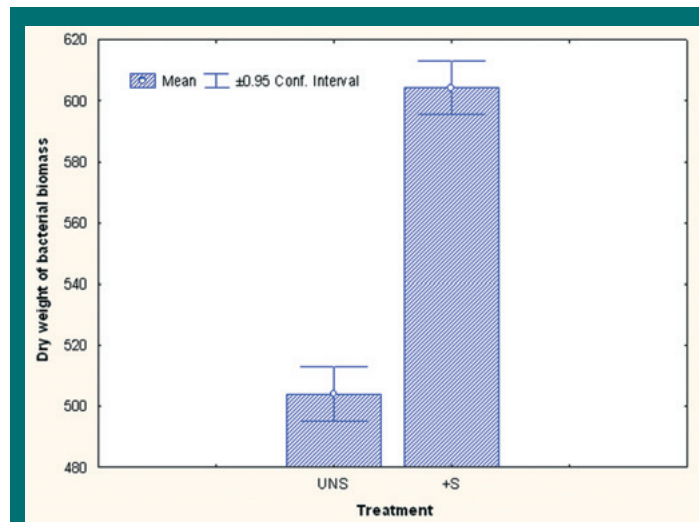


Fig. 1. Mean bacterial biomass (with 95% confidence intervals) using unsupplemented (UNS) and supplemented (+S) media (t-test for comparing the means of both groups, $t = 16.59$, $P < 0.001$)

novel compound. Both vessels were allowed to equilibrate and were subject to identical environmental / incubation conditions. The vessels were then inoculated with a culture of Bacterium x at an equal culture density and the fermentation allowed to proceed until all the available nutrients had been exhausted and bacterial growth had ceased. The entire volume of culture media in each fermentation vessel was then removed and filtered to recover the bacterial biomass. This experiment was repeated 25 times and the biomass measurements are given in Statnote 2.

How are the calculations done?

To determine whether an individual measurement is typical of a population requires knowledge of how individual measurements vary; i.e., the SD of the population (Statnote 2). Similarly, to determine the degree of error associated with a sample mean requires knowledge of how means vary; i.e., the SE of the mean (Statnote 2). Extrapolating this principle further, to determine whether there is a significant

difference between the means of two samples, knowledge is required of how the differences between two sample means would vary. Hence, for each of our two samples, the mean is calculated and the difference between the mean of the unsupplemented (U*) group subtracted from the mean of the supplemented (S*) group. This difference represents the treatment effect of the experiment; i.e., the degree to which the media supplement may have increased bacterial biomass. If this experiment was repeated many times, several estimates of $U^* - S^*$ would be obtained and we could construct a frequency distribution of the differences between the means. However, if the distribution of the means from the supplemented and unsupplemented groups are normally distributed, then the distribution of the differences between pairs of means taken from these two populations will also be normally distributed. As a result, the standard normal distribution (SND) can be used to test whether there is a true difference between the means.

The group means of the

two treatments differ by 100.16 units (Fig 1) and the two samples exhibit little overlap suggesting a real effect of the supplement (Fig 2). There is, however, variation in microbial biomass between replicate flasks within each group. Hence, is the difference between the means actually attributable to the effect of the supplement or could it be accounted for by random variation between the flasks? To decide between these two alternatives, the treatment effect ($U^* - S^*$) is compared with the degree of variation pooled from both groups by carrying out a 't' test. The statistic 't' is the ratio of the difference between the two means to the SE of the difference between the means:

$$t = \frac{U^* - S^*}{\sqrt{s^2 (1/n_1 + 1/n_2)}}$$

where 's²' is the average variance of both groups and n_1 and n_2 are the number of observations within each group ($n_1 = n_2 = 25$). Note that this calculation is similar to that carried out in Statnote 2 when an individual value 'x' was converted so that it became a member of the SND. In the present example, the 't' distribution is used instead of the SND because 't' describes the variation of means calculated from small numbers of observations more accurately. Hence, when 't' is calculated, the difference between the means ($U^* - S^*$) becomes a member of the 't' distribution. The probability of obtaining a value of 't' of this magnitude by chance from two groups of samples when each comprises 25 observations is obtained either from statistical software or by consulting statistical tables (Fisher and Yates, 1963). In the present example, a value of $t = 16.59$ was obtained and is taken to a table of the 't' distribution, entering the table for $n_1 + n_2 - 2$ DF (DF = 48). When 't' is equal to or

greater than 2.01 (the value at $P = 0.05$ for 48 DF), the value is in a region of the distribution which includes the 5% most extreme values. Hence, $t = 16.59$ is an unlikely value to occur by chance and therefore, there is a real difference between the two means.

One-tail and two-tail tests

It is possible to propose two different null hypotheses. First, that the addition of the supplement (S) would have no effect on bacterial biomass. This hypothesis does not specify whether a significant increase or a decrease in biomass would be necessary to refute the null hypothesis. In this case, a two-tailed test would be appropriate, i.e., both tails of the 't' distribution are used to test the hypothesis. Second, that the supplement would only increase biomass since it may be known in advance that it could not significantly decrease biomass. If the hypothesis specifies whether a positive or a negative effect is necessary to refute the hypothesis, a one-tail test would be appropriate. Some statistical tables indicate both the one-tail and two-tail probabilities corresponding to particular columns. Most statistical tables, with some notable exceptions, only indicate the two-tail probabilities. To find the one-tail probabilities in a two-tail table, halve the probabilities; i.e., the 5% one-tail probabilities are found in the 10% two-tail column.

Paired and unpaired 't' tests

An experiment involving two treatments or groups can be carried out in two different ways, *viz.*, the unpaired and the paired methods. The experiment described in our scenario was carried out using an unpaired design; *i.e.*, the

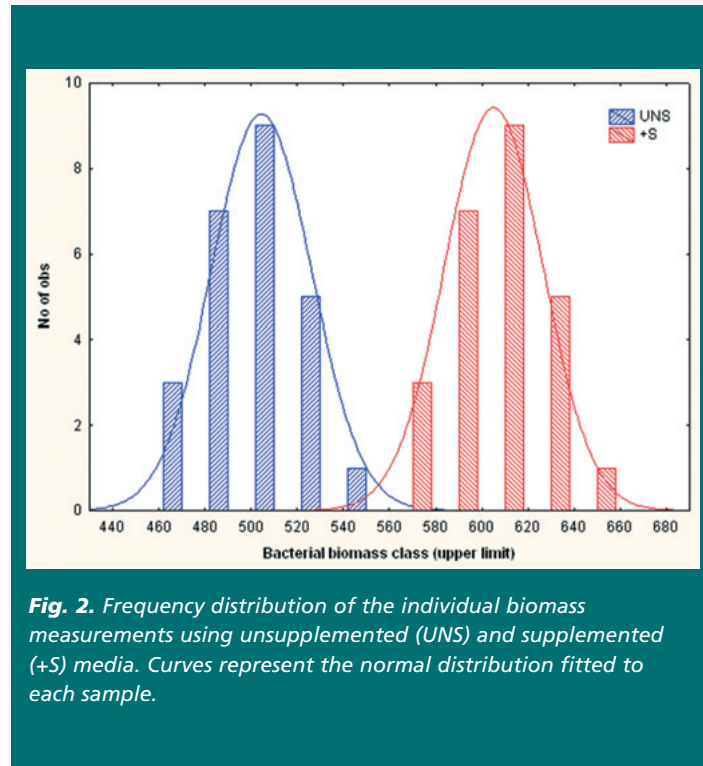


Fig. 2. Frequency distribution of the individual biomass measurements using unsupplemented (UNS) and supplemented (+S) media. Curves represent the normal distribution fitted to each sample.

media supplement was allocated at random and without restriction to half of the 50 original flasks. In a paired design, however, the 50 flasks are first divided into 25 pairs with the intention of processing a single pair (one supplemented, the other unsupplemented) on each of 25 days and second, the supplement is then allocated to one flask of each pair independently and at random. Hence, there is a restriction in the allocation of the treatments to the flasks and a different analysis is required. In a paired design, the 't' test is made as follows:

$$t = d^* / (sd / \sqrt{n})$$

In this case, 'd*' is the mean of the differences between each of the 25 pairs of observations and 'sd' is the standard deviation of these differences. The same 't' table is used for determining the significance of 't'. In a paired 't' test, however, a different rule is used for entering the 't' table; *viz.*, 't' has n-1 degrees of freedom, where 'n' is the number of pairs of subjects. Again, one-tail or two-tail tests

may be made as appropriate.

Advantages of pairing

Is a paired or an unpaired design the best method of carrying out the experiment? Each type of design has advantages and disadvantages. A paired design is often employed to reduce the effect of the natural variation that always exists between flasks or replicates. How this may be achieved can be seen by examination of the formula for the unpaired 't' test. A value of 't' is the difference between the two treatment means divided by the SE of this difference. If variation among flasks is large, say from processing them at different times of the day or on different days, it will increase the SE of the difference and lower the value of 't' even if the difference between means is large. Notice, however, that in an unpaired design, the 't' table is entered for 48 DF. Pairing the flasks may reduce the SE because paired 't' is calculated from the differences between pairs of observations. In other words,

the effect of the experimental treatment is being determined within a matched pair of flasks. Pairing should only be considered, however, if there is evidence that it actually reduces the variability, e.g., pairing supplemented and unsupplemented flasks on the same day when the day of assay does significantly affect the measurement. If there is no reduction in the SE by pairing, *i.e.*, it does not matter which day the samples are measured, then there is a disadvantage of the paired design because the 't' table is entered with only 24 DF (1 less than the number of groups). Entering the 't' table with a smaller number of degrees of freedom means that a larger value of 't' will be required to demonstrate a significant difference between the means.

The 't' test is a useful method of comparing two groups when the data approximate to a normal distribution. In many cases in microbiology, however, data may not conform to this distribution and different methods may be required to test differences between the groups and these methods will be reviewed in Statnote 4.

References

- Fisher RA & Yates F (1963) Statistical Tables for Biological, Agricultural, and Medical Research. Sixth Ed, Longman, Bristol.
- Hilton A & Armstrong RA (2005) Are the data normal? *Microbiologist* June 2005, **Vol 6** No 2, 34-35
- Hilton A & Armstrong RA (2005) Describing the normal. *Microbiologist* September 2005 **Vol 6** No 3, 30-32

Dr Anthony* Hilton and Dr Richard Armstrong**
*Pharmaceutical Sciences and **Vision Sciences, Aston University, Birmingham, UK