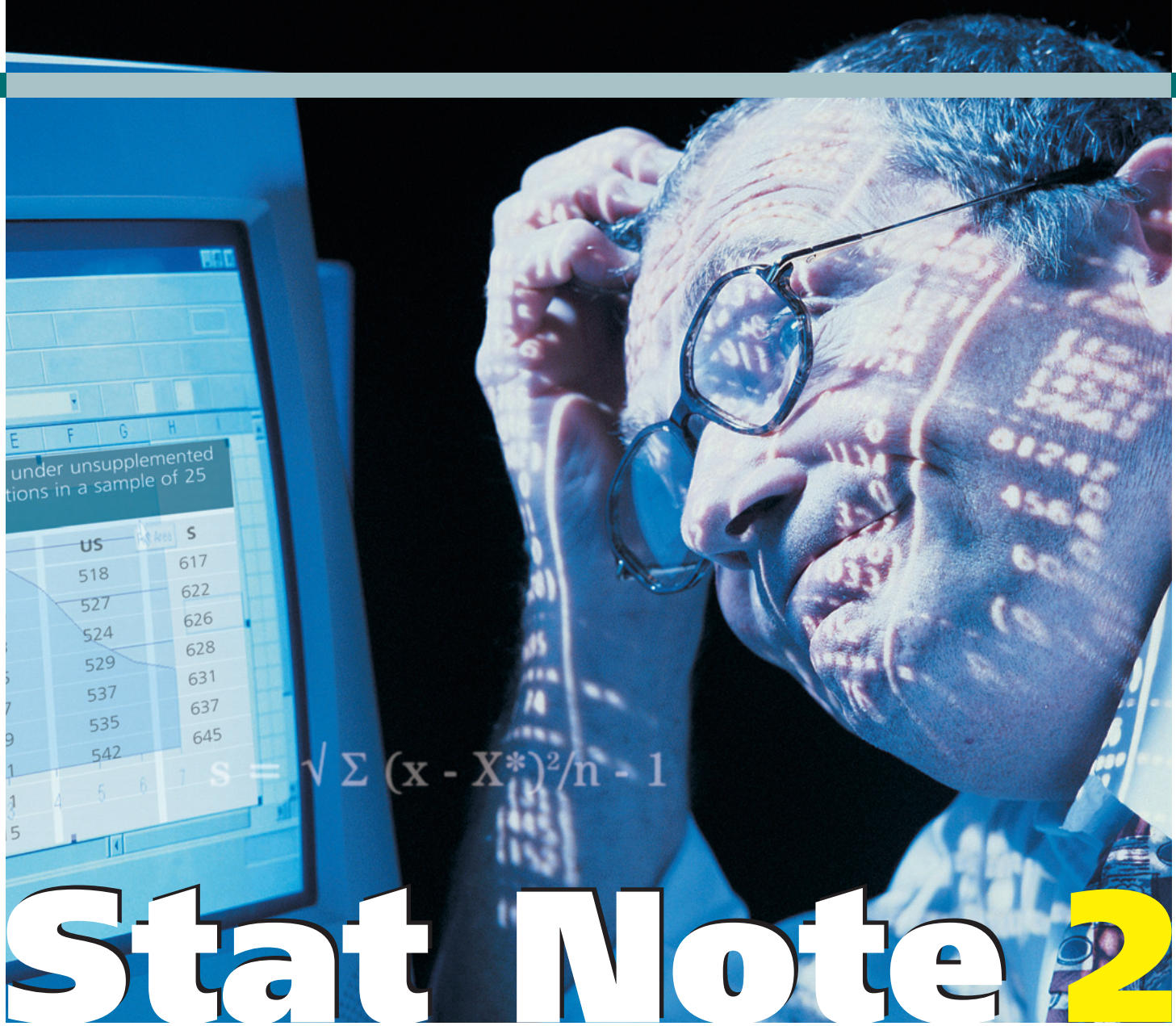

Stat Note 2

In the second of a series of articles about statistics for biologists, **Anthony Hilton** and **Richard Armstrong** describe the application of normal distribution to some common statistical problems

In our first Statnote (*Microbiologist*, June 2005) we described two procedures involving chi-square (χ^2) and the Kolmogorov-Smirnov (KS) test to determine whether a sample of data can be considered to come from a normal distribution. If the sample does not come from a normal distribution then a number of statistics can be calculated that describe the central tendency (mean) and degree of spread of the sample. In addition, we can use our sample of measurements to make inferences about the mean and spread of the population from which the sample has been drawn. This Statnote describes the application of the normal distribution to some common statistical problems including how to determine whether an individual observation is a typical member of a population and how to obtain the *confidence interval* for a sample mean.

The scenario

A hypothetical experiment was carried out to investigate the efficacy of a novel media supplement in promoting the development of cell biomass. Two 10-litre fermentation vessels were sterilised and filled with identical growth media with the exception that the media in one of the vessels was supplemented with 10ml of the novel compound under investigation. 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 growth had ceased. The entire volume of culture media in each fermentation vessel was then removed and filtered to recover the bacterial biomass which was subsequently dried and the dry weight of cells measured. This experiment was repeated 25 times and the dry weight of biomass produced in each experiment recorded in the table at right (Table 1). This Statnote will only be concerned with analysis of the data from the supplemented culture and in Statnote 3 (*Microbiologist*, December 2005) the same scenario will be used to describe how to determine the significance of any difference between media with and without supplement.

How are the calculations done?

Describing the normal

If our sample of measurements of bacterial biomass (N = 25) on supplemented media (X) is plotted as a *frequency distribution* (Fig. 1), the measurements appear to be more or less symmetrically distributed around a central tendency or average value. If the number of measurements were to be increased to a large number and the class intervals

of the distribution reduced to zero, the data would approximate closely to a bell-shaped curve called the normal distribution (also known as a Gaussian distribution). Many measurements in the biosciences follow this type of distribution. In the present case, the sample data did not deviate significantly from normal as indicated by a KS test (see Statnote 1).

The normal distribution can be described by two statistics:

(a) The average or *arithmetic mean* of the population ($\mu = \Sigma x/n$) where 'x' stands for each item in the sample taken successively. Note that the mean of a sample of measurements taken from this population is designated as 'X*' or 'x*'.

(b) The *standard deviation* (SD) of the population, i.e., the distance from the mean to the point of maximum slope on the curve ($SD = \sqrt{\Sigma(x - \mu)^2/n}$). Hence, the SD describes how close the data cluster around the mean. Note that the SD of a population is given the symbol σ while that of a sample is often designated as 's' or 'σ_{n-1}'.

To calculate the SD we need to know 'μ', the mean of the population. However, in most circumstances we wish to calculate the SD of a small sample of measurements taken from a much larger population. In this case, we do not know the exact value of 'μ' but we can calculate the sample mean 'X*'. Hence, to calculate the SD of a sample of measurements we can use the formula for the SD defined above but with three changes:

(a) The SD of the population 'σ' is replaced by the symbol 's', the SD of the sample.

(b) μ is replaced by X*, the mean of the sample.

(c) 'n' is replaced by 'n-1', a quantity called the *degrees of freedom* (DF).

The calculation of the SD involves the subtraction of individual observations from their mean, which are then squared and summed. However, if there are 'n' observations, once 'n - 1' observations have been subtracted from the mean, we can immediately calculate the last deviation because the sum of all of the deviations from the mean would be zero. In other words, 'n' observations only provide 'n - 1' *independent* estimates of the deviations from the mean. As a general rule, the DF of a statistical quantity is the number of observations making up that quantity minus the number of parameters that have to be calculated from the data to obtain that quantity.

Hence, the formula for the SD of a sample is:

$s = \sqrt{\Sigma (x - X^*)^2/n - 1}$.

If several estimates of the same quantity are made in a study, it is common practice to report the mean and the SD of the sample. In the present example, we would describe our sample of biomass measurements on supplemented media as having a mean of 604.28 and an SD of 21.16.

Another useful way of expressing the variability of a sample is as the *coefficient of variation* (CV) defined as the SD expressed as a percentage of the mean:

$CV = s \times 100/x^*$

The CV provides a *standardised* method of expressing the variability of a measurement in an experiment. Different variables in the biosciences often have characteristic CVs that are stable across experiments, so it may be possible to obtain an estimate of the variability of a quantity in advance by examining the results of previous experiments. The CV is therefore useful in planning experiments. In the present case, the CV for the supplemented data is 3.5%.

Table 1. Dry weight of bacterial biomass under unsupplemented (US) and supplemented (S) growth conditions in a sample of 25 fermentation vessels.

US	S	US	S	US	S
461	562	506	607	518	617
472	573	502	600	527	622
473	574	501	603	524	626
481	581	505	605	529	628
482	582	508	607	537	631
482	586	500	609	535	637
494	591	513	611	542	645
493	592	512	611		
495	592	511	615		

The equation of the normal distribution

The mathematical equation that describes the normal distribution is given as follows:

$y = 1/\sigma\sqrt{2\pi}\ (e^{-(x-\mu)^2/2\sigma^2})$ (Snedecor & Cochran 1980)

This equation enables the height of the normal curve (y), to be calculated for each individual value of 'x' providing that 'μ' and 'σ' are known. This equation also enables the proportion of observations that fall a given distance from the mean to be calculated. In any normal distribution, approximately 68% of the observations will fall one SD above and below the mean. Hence, the probability is 68% or P=0.68 that a single measurement from a normal distribution will fall between these limits. Similarly, the probability is P=0.95 that a single measurement will fall approximately two SD above and below the mean. Each type of variable has a characteristic normal distribution of values with a typical mean and standard deviation. However, statistical tables of the normal distribution, called 'z' tables, have been calculated for a distribution termed *'the standard normal distribution.'* If we wish to use these tables in statistical tests, then we have to convert our measurements so that they are members of the standard normal distribution.

Is a single observation typical of the population?

The standard normal distribution has a mean of zero ($\mu = 0$) and a standard deviation of one unit ($\sigma = 1$) and provides the basis of many useful tests. For example, it may be important to determine whether a single observation 'x' is a typical or atypical of a population of measurements. To make this test, the original observation 'x' has to be converted so that it becomes a member of the standard normal distribution 'z':

$z = \pm (x - \mu)/\sigma$

Tables of the standard normal distribution can then be used to determine where 'z' is located relative to the mean of the distribution, i.e., does it fall near the mean of the distribution (a typical value) or out in one of the tails of the distribution (an atypical value).

An important question is how atypical does 'z' have to be before we would consider it not to be a member of the population? By convention, we will consider 'x' to be a typical member of the population unless it is located in the tails of the distribution which include the 5% most extreme values. The value of 'z' that separates

the typical values (95% of the distribution) from the atypical values (5% of the distribution) is actually 1.96. Hence, if our calculated value of 'z' is equal to or greater than 1.96, we would consider the measurement to be atypical.

As an example, assume that we make an additional estimate (x) of bacterial biomass under supplemented conditions and obtain a value of 550. Is this value typical of the 'population' of values defined in Fig 1. Subtract the mean from 'x' and divide by the SD to convert 'x' to 'z'. A value of z = -2.56 was obtained which is greater than 1.96, the value of 'z' that cuts off the 5% most extreme observations in the population. Hence, 550 is not typical of the values obtained previously and there would be some doubt as to whether the conditions of the original experiment had been exactly reproduced in making our additional estimate of 'x'.

The variation of sample means

If we repeated the study on supplemented media with several samples of 25 we would not necessarily obtain the same mean value each time, i.e., the means of samples also exhibit variability. In this case, we might want to know how good an estimate our individual sample mean was of the population mean. To answer this question requires knowledge of how means from a normal distribution of individual measurements themselves vary. To understand this concept, it is necessary to quote an important statistical result termed *'The Central Limit Theorem.'* This states that means from a normal distribution of individual values are themselves normally distributed with mean 'μ' and SD s/√n, where 'n' is the number of observations in the sample. In addition, the means of many non-normal distributions will be normally distributed as long as the samples are large enough. It is important to distinguish the quantity s/√n, the SD of the population of sample means or *'standard error of the mean'* from 'σ' or 's' the SD of a population or sample of individual measurements.

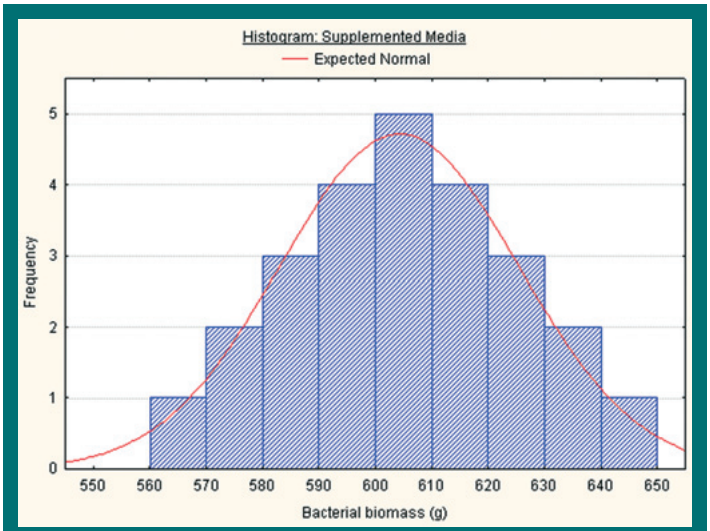

Figure 1. *Frequency distribution of bacterial biomass on supplemented media. Fitting the normal distribution KS = 0.079, P > 0.05).*

How to fit confidence intervals to a sample mean

The standard error of the mean is often plotted on a graph as a *confidence interval* or error bar, and indicates the degree of confidence that we have in our sample mean as an estimate of the population mean. Confidence intervals are calculated as follows:

(a) If a single observation 'x' comes from a normal distribution then the probability is 95% (P = 0.95) that 'x' will be located somewhere in the distribution between $\mu \pm 1.96\sigma$.

(b) Similarly, if a sample mean X* comes from a normal population of sample means then P = 0.95, that X* lies between $\mu \pm 1.96\ \sigma/\sqrt{n}$.

(c) Hence, we can write P = 0.95 that μ lies between X* ± 1.96 σ/√n.

There are two problems with this approach. First, in the majority of studies, the sample mean X* is based on a small sample of measurements. Hence, we do not know the value of 'σ' only the SD of the sample 's'. Hence, we substitute 's' for 'σ'. Second, we cannot be certain about the exact shape of the distribution and therefore whether the value of Z = 1.96 is accurate enough to judge whether a sample mean is atypical of the population. Instead, we use a different value that more accurately describes the behavior of small samples, *viz.*, a value from a related distribution called the 't' distribution. The 't' distribution will be discussed in more detail in the next Statnote.

(d) Hence, the 95% confidence interval (CI) of a sample mean is given as CI = X* ± 't' (P = 0.05, DF = n - 1) s/√n. For our supplemented biomass data, the 95% CI were estimated to be 604.28 ± 8.72.

Therefore, we are 95% confident that the population mean will fall between the calculated limits. The 95% confidence intervals are often plotted as error bars. It is important to be clear what the error bar represents since investigators may plot the SD of a sample, the standard error of the sample mean, or the 95% confidence intervals and each conveys different information. In addition, error bars must not be used to make judgments as to whether there are significant differences between two or more means on a graph. The confidence intervals of two sample means are calculated using the standard errors appropriate to those sample means alone. To test whether the two means are different requires another form of standard error, i.e., the *'standard error of the difference between two means'*, and this will be discussed in Statnote 3.

Conclusions

If a sample of measurements comes from a population that is normally distributed, we can use several statistics to describe our sample, such as the mean, SD, and CV. In addition, we can determine how atypical an individual measurement has to be before we would consider it not to be a member of a specific population.

Furthermore, we can use our sample to make inferences about the population from which the sample is drawn including making estimates of the population mean and fitting confidence intervals to a sample mean.

Reference

■ Snedecor GW and Cochran WG (1980) *Statistical methods,* 7th Ed. Iowa State University Press, Ames Iowa.

Dr Richard Armstrong and Dr Anthony Hilton
Life and Health Sciences, Aston University