The Importance of Statistics and Replication in Microbiology

Granda: I play much better if I have an energy drink at half-time!

Photo by Kelly L from Pexels

James I Prosser

School of Biological Sciences, University of Aberdeen, United Kingdom:

The importance of statistics and replication in microbiology

Storyline

Scientific studies aim to explain observed phenomena through hypotheses that generate prediction which can be tested by comparison with experimental data. Both initial observation of phenomena and experimental testing involve collection and analysis of data. While these data may be qualitative, both the description of natural phenomena and the critical testing of hypotheses are more thorough, exacting and complete if quantitative data are employed. For example, we might observe that pasteurisation reduces bacterial abundance in milk, but it is of much greater value if we observe that pasteurisation reduces abundance to 50% or 1% or 0.0001%. Quantitative data are therefore fundamental to microbiology and statistics provides with the techniques for organisation, presentation, analysis and interpretation of these data. Statistics can be of two types. *Descriptive statistics* summarise data, visually or, usually, numerically. In contrast, i*nferential statistics* are used to interpret and draw conclusions from our data. Some studies do not require statistical analysis, e.g. if they involve only qualitative data. In practise, most generate quantitative data and require statistical analysis. It is therefore essential for any microbiologist to understand the principles underlying statistical analysis and to identify the statistical methods required for collecting or analysing data.

The discussion below outlines some of the basic principles of statistics as applied to continuous data in relatively simple situations. It is not possible to discuss the wide range of microbiological data encountered, complex experimental design, or all of the assumptions on which the analyses are based, but the principles apply throughout. Statistical analysis necessarily involves mathematics and equations and these are presented in boxes. In the past, these calculations were performed on calculators or spreadsheets, but now routinely involve statistical software packages. While this is convenient, often necessary and generates more detailed information, it does introduce the real danger of accepting software output without understanding the underlying principles and can lead to mis-interpretation.

The Microbiology and Societal Context

Statistical analysis is required for any microbiological study that involves quantitative data. It therefore cuts across, and is required in all aspects of microbiology and SDGs.

1. General comments

a. *Populations and samples*. In analysing data, we distinguish between a population, which is the largest collection of entities for which we have an interest, e.g. all bacteria in a culture, or several cultures grown under different conditions. It is rarely feasible to measure properties of each member of a population and we usually measure properties of a portion or sample of the population, from which we infer properties of the population. In doing this it is important to sample populations randomly to avoid bias i.e. each member of the population should have an equal chance of being chosen. Precision and, consequently, the amount of information we obtain, increase as the number of samples and sample size increase.

b. *Variability and error*. Statistical analysis is required because of variability, e.g. bacterial abundance in pasteurised milk from different shops or in different cartons will differ. Some of this variability will be *systematic* (potentially explainable), while some will be random or *experimental error*. The latter arises because of inherent variability in the experimental material or lack of uniformity in physical conduct of the experiment. Both types of error must be minimised to improve the power of any statistical tests, e.g. by handling experimental material

to reduce the effects of inherent variability, refining experimental technique or common sense.

c. *Replication.* One technique for reducing error is replication. If a treatment appears more than once in an experiment it is said to be replicated. Replication serves two major functions. First, it provides an estimate of experimental error, which is required for tests of significance and confidence limits. If there is just one treatment i.e. a single replicate, there is no information about experimental error and it is impossible to determine whether differences between this treatment and another are due to differences in treatments or to differences between experimental units. Second, replication improves the precision of an experiment by reducing the standard deviation of the treatment mean (see below).

The number of replicates required depends on the precision required, which may be difficult to decide in advance, but it is pointless performing an experiment that will not give the required precision. There are other sophisticated ways of reducing experimental error but most are just common sense. Elimination of careless technique is essential, as this is often non-random and biased and constitutes inaccuracy rather than variability.

d. Precision and accuracy. Precision and accuracy are considered synonymous colloquially, but have different meanings in statistical analysis. Both are measures of error but accuracy describes how close observations are to their 'true' value, while precision describes how close measurements are to each other.

2. Descriptive statistics

a. *Averages.* Averages or measures of central tendency are important ways of describing data. For example, if we measure the length of 100 cells, we could present the data as a list or a histogram, but it is much more convenient to describe it as an average. The most common measure of the average is the mean (Box 1), which is easy to calculate, stable to fluctuations in sampling and capable of algebraic manipulation. Another average is the median, the central value when all measurements are ordered, which is more stable with respect to extreme values.

b. *Variability*. A more complete description of data requires a measure of scatter or variability, the two most common being the variance and the standard deviation. We could quantify variability by summing the difference between each value and the mean, but, by definition, this sum would = 0. The squares of each difference will be positive and the sum of these squares is called the total corrected sum of squares (*SS*). *SS* increases with population or sample size, and so must be standardised to compare different sized populations. For populations, this is achieved by dividing *SS* by the population size, *N*, but for sample means we divide by $n - 1$, where n is the sample size (Box 1). This reduces bias and overconfidence in estimating variability in the population from that in a sample. This bias results from the fact that we have already calculated total variability when calculating the mean and can therefore calculate the n^{th} difference once we know the other $n - 1$ values. Therefore we say that this final value is 'not free to move' and that we have $n - 1$ *degrees of freedom*. The values we obtain are called the population or sample variance, but its units are the square of our measurements, e.g. cm^2, which is not very intuitive, and it is more usual to describe variability as the *standard deviation*, which is the square root of the variance (see Box 1). The standard deviation can be viewed as the average distance of values from the mean.

c. *Correlation coefficient.* The statistics above are 'univariate', involving a single variable, e.g. height. We often measure two or more variables in the same individual, e.g. height, weight, age, which are described by bivariate or multivariate statistics, respectively. Associations between two variables, x and y, can be visualised by plotting them for each individual on the x and y axes. The strength of the association can be quantified by correlation analysis and calculation of a correlation coefficient. One example is the Pearson's r or product-moment correlation (see Box 1).

Equations for correlation coefficients are more complicated (see Box 1), but the numerator represents covariance, a measure of how x varies with y, calculated as the sum of the product of differences between each variate and its mean. This is equivalent to variance of univariate data, while the denominator normalises this with respect to the standard deviations of each variate.

The correlation coefficient quantifies the *strength* of the association between two variates and has no units, as it represents the ratio of variances. It has values between -1 (complete negative association), through 0 (no association) to +1 (complete positive association). The closer to -1 or +1, the stronger the association. It not a measure of quantitative change of x with respect to y. Most importantly, it gives no information on cause and effect. Thus, a high correlation coefficient does not mean that one variate is affecting, or is affected by the other.

d. *Linear regression*. Linear regression quantifies the relationship between a dependent variable y and an independent variable x. In this situation we have control of x and assume that there is a linear relationship between y and x, e.g. the relationship between optical density (y) and concentration (x). Linear regression estimates the strength of the relationship and enables calculation of y for any value of x, but only within the range of x and y values that you have measured. The relationship between x and y is represented by

$$y = a + bx + E$$

where a and b are the intercept and slope of the line that minimise the error (E), i.e. they give line of best fit. Equations for linear regression are given in Box 1 and statistical software provide other statistics on the quality of fit, confidence in a and b, etc.

3. *Inferential statistics*

a. Distributions. The mean and variance are useful in describing the 'most likely event' and the variation around it, but they are of more statistical use if we know how values are distributed. Many continuous biological characters follow the Gaussian or Normal distribution (Fig. 1).

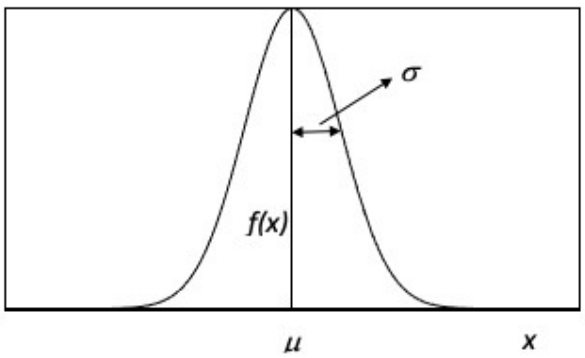

Fig. 1. Normal distribution of *f(x)*, the frequency of a variable *x*, as a function of *x*.

Thus if *x* represents height of people, and *f(x)* the frequency of occurrence of different heights, the plot of *f(x)* vs *x* will give a bell-shaped curve with mean, μ, and standard deviation, σ. Note

that the variation is continuous and distributed equally about the mean, i.e. the distribution is symmetrical and deviations from it are equally likely in each direction. In theory there are no upper or lower limits to *x* but the frequency of extremes is very low. The standard deviation measures the distance from the mean to the point of inflexion on the curve. (It must be remembered, however, that not all things are distributed normally.)

The normal distribution can be used to calculate the proportion of the population with characteristics above or below a particular value. However, these calculations are not straightforward without the use of computers and depend on μ and σ. Before computer software was readily available, this problem was solved using the standard normal distribution or *z* distribution. If we subtract the population mean from each individual value, we would have a mean of 0. Similarly, as the standard deviation is the average deviation from the mean, division of each difference by the standard deviation will give a new standard deviation of 1. Standardisation is therefore achieved using the equation: $z=\frac{x-\mu}{\sigma}$. The *z*-distribution has the same shape as the normal distribution, but with a mean of 0 and a standard deviation of 1. In the past, this enabled calculation of different probabilities (areas under the curve) using tables of the standard normal distribution, and this transformation now makes computation quicker. Both normal and *z*-distributions can be described by an equation (see Box 1) and approximately 68% of values fall within the mean ± standard deviation, while 95% and 99.7% fall within 2 and 3 standard deviations of the mean (Fig 2).

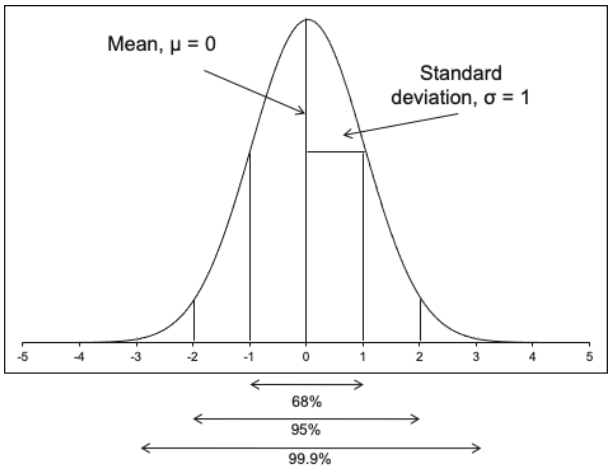

Fig. 2. Standardised normal or *z*-distribution, indicating the probabilities (areas under the curve) within the mean ± 1, 2 and 3 standard deviations.

b. Statistical significance. Inferential statistics enables us to draw conclusions from our data. For example, we may want to know if bacteria grow differently on different media or at different temperatures. Effectively this is hypothesis testing, in that we are testing the hypothesis that the medium or temperature affect growth. Any test of an idea or hypothesis, either from theoretical reasoning or suggested by results from earlier experiments, should involve a clear statement of objectives and we define the Null Hypothesis, which is given the symbol H_o. Statistically, H_o is a statement of 'no difference' between two sampled populations that may be expected to differ, e.g. through different treatments.

Significance tests are the basis of inferential statistical analysis and will be discussed initially by imagining the situation where we have sampled two different populations, X and Y, that we suspect to be different. We are therefore testing the null hypothesis, H_o, that the population means are equal, $\mu_1 = \mu_2$, but we can perform the test by placing emphasis on an alternative hypothesis, H_A, $\mu_1 > \mu_2$, ignoring any evidence that $\mu_1 < \mu_2$. We take samples from each population and calculate sample means, $\overline{x}_1$ and $\overline{x}_2$. If $\overline{x}_1 > \overline{x}_2$, then we might reject H_o in

favour of H_A. However, this result may be because μ_1 really is greater than μ_2 or because it just so happens that the samples taken made it look like that, because of the uncontrolled sources of error and variation. This is always a risk when sampling. The smaller the sample and the smaller the difference between μ_1 and μ_2, the greater the risk.

To distinguish between these reasons and to choose H_o or H_A we must introduce the concept of significance. To do this we calculate the probability that a difference at least as extreme as $\bar{x}_1 - \bar{x}_2$ (the observed difference) arose by chance on the assumption that H_o is true (i.e. $\mu_1 = \mu_2$. If the probability is small, then we may conclude with reasonable (but not absolute) certainty that the difference is real. If the probability is large then we conclude that the difference is not real but arose from chance as a result of sampling and experimentation.

c. Standard error and confidence intervals. We must now consider our confidence that the sample mean accurately represents the population mean. We could take many samples, each giving a different mean, and it can be shown that the means of all of these samples is distributed normally and that the standard deviation of the mean, also called the standard error, is given by the equation $s_{\bar{x}} = s/\sqrt{n}$. This equation defines, quantitatively, how replication increases our confidence that the sample mean reflects the population mean, with variance decreasing, and accuracy increasing, in proportion to the square root of the sample size.

Data are often presented as the mean ± standard deviation or standard error. An alternative is to use confidence limits which will contain a parameter with a probability of 95% (or some other value). These are called the 95% confidence limits and, for large sample sizes, can be determined using the standardised normal distribution, based on our sample mean and standard deviation: $z = (x - \bar{x})/s$. Now, we want to know the value of x that will give a probability of 2.5%. This value is 1.96 (see Fig. 2), so the equation becomes: $x = \bar{x} \pm 1.96\, s$. If the sample size is relatively small (<40) we would use the equivalent value from the *t*-distribution (see below) rather than the normal distribution.

d. *Student's t-test*. Suppose we want to test a theory that predicts that the diameter of a fungal hypha, under certain conditions, will be 11 µm, i.e. that the population mean, µ = 11 µm. In statistical terms we are testing the null hypothesis H_o: $\mu = \mu_o = 11$ and we measure the diameter of 100 hyphae and calculate a sample mean of 11.14 and a standard error of 0.11 µm. We then transform our variable to the standardised normal distribution, i.e. if H_o is true, then $z = (\bar{x} - 11)/0.11$ is normally distributed with mean 0 and variance 1.

We assess the significance of the observed value $\bar{x}$ = 10.86 by calculating the probability that a mean value at least as extreme as 10.86 can occur by chance, assuming H_o is true. Any value less than 10.86 is certainly more extreme than the observed $\bar{x}$, but so also is any value greater than 11.14, i.e. we must consider extremities above or below the mean and determine the probability that $\bar{x} < 10.86$ and $\bar{x} > 11.14$. This probability (calculated from tables of *z*-distribution or from computer software) is 2 x 0.102 = 0.204. This means that if H_o is true we would expect a sample mean of 10.86 from approximately 20%, one sample in five. We therefore cannot conclude that H_o is false. i.e. µ = 11 with a probability of 1 in 5. By convention, we start doubting H_o when the probability reaches approximately 1 in 20, or 5% (0.05). This is sometimes described as the 5% or 0.05 level of significance.

No null hypothesis may be considered in isolation. There is always an alternative hypothesis, H_A, even if it is not stated explicitly. Above we tested H_o:$\mu = \mu_o$ against H_A:$\mu \neq \mu_o$ i.e. $\mu < \mu_o$. Therefore, when we considered values of $\bar{x}$ at least as extreme as the observed $\bar{x}$ = 10.86 this included any value < 10.86 but also any value > 11.14. This is called a two-tailed test. If we had declared H_A: $\mu < \mu_o$ then any $\bar{x} > 11$ would not have been extreme, since it would have been more likely to come from a population with $\mu = 11$ than with $\mu < 11$. So in this case we would use a one tailed test and would have obtained a probability of 0.102.

Another consideration is sample size. With large sample sizes, n > 30-40, $\bar{x}$ and s are reasonably accurate estimates of μ and σ. However, for smaller sample sizes, we cannot have confidence in the accuracy of s^2 and, rather than using the z distribution, we use the Student's t-distribution. We perform a similar transformation: $t = (\bar{x} - \mu_o)/s_{\bar{x}}$ and assume that t follows the t-distribution with n -1 degrees.

Note that in both of these examples, the statistic being calculated contains the 'difference' being investigated, while the denominator is a measure of experimental error. In other words we are determining the ratio of an observed difference due to a potential effect as a proportion of the experimental error. This approach applies in the possibly more common situation of comparing two sample means, rather than comparing one mean with a fixed value. In this case, the numerator is the difference between the sample means, while the denominator is a measure of the combined experimental error. For a small sample size and different sample sizes, the equation estimation of common variance becomes more complicated (Box 1).

e. *Analysis of variance*. In many experiments we want to compare several means, rather than just two. For example, we might be investigating the effect of 5 different growth media on biomass yield, with a null hypothesis that medium has no effect on biomass. We have 4 replicates for each medium, giving a total of 20 biomass measurements, and calculate 5 mean values, each from 4 replicates. To assess whether growth medium affects yield, we can perform an analysis of variance.

The first step is to assess the mean and variance across all 20 cultures, the latter being the total corrected sum of squares (see above). The analysis of variance splits this total variance into a number of component parts which we believe to be related to different causal circumstances (treatments or factors), in our case use of different media, and experimental error. It calculates the variances about the means of these components and assesses the significance of these variances.

In our case we have two sources of variation, treatment (growth media) and inherent variation (experimental error). Experimental error can be estimated by calculating, independently, the sum of squares within each treatment and then summing these values to give the 'within sum of squares, SS_{wit}. If growth medium affects biomass, then SS_{wit} will be less than SS_{tot}, and the remaining variance due to the treatment, i.e. between treatment variance or SS_{bet}.

There will always be some random variation but if this is larger than we would expect then there may be no real difference between treatments. We therefore draw up an Analysis of Variance (AOV) Table (Box 2) which contains the sum of squares. It also contains the within and between mean squares, which are the sum of squares divided by the respective number of degrees of freedom. If H_o is true, and there is no treatment effect, we would expect the MS values to be approximately equal as both will estimate total variance. We therefore calculate the ratio of MS values (the variance ratio) and either compare with tabulated values of the *F*-distribution for a probability of 0.05 with the appropriate degrees of freedom, or use statistics software to generate a *p* value. F-distribution values are based on degrees of freedom associated with both between and within variation.

Pupil participation

1. Class discussion of the importance of replication and statistical analysis underpinning scientific discoveries reported in the news.

2. Exercises

a. Basic descriptive statistics can be calculated for pupil characteristics, e.g. height, weight. The effect of sample size and the standardised normal distribution can be illustrated by measuring height in samples of different size and correlation coefficients can be calculated for height and age. Inferential statistics can be illustrated by comparison of means of different groups within a class, or between different classes.

b. The covid pandemic has highlighted the need for quantitative microbiological data and limitations in basic understanding of these data and the statistical analyses employed. Published covid statistical data can therefore be used to illustrate the basic principles of statistics and ways in which data can be interpreted, and misinterpreted.

Further reading

There are many introductory books on statistics, including those for those studying biology and microbiology, and equally good web sites, a good example being https://www.scribbr.com/category/statistics/.

Box 1. Statistical equations referred to in text.

Population mean $\mu = \frac{\sum X}{N}$, where *N* is population size and X is individual measurement Sample mean $\bar{x} = \frac{\sum x}{n}$ where *n* is population size and *x* is individual measurement
Variability: Population: Total corrected sum of squares ¿ $\sum (X-\mu)^2$, variance $\sigma^2 = \frac{\sum (X-\mu)^2}{N}$, standard deviation $\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$ Sample: Total corrected sum of squares ¿ $\sum (x-\bar{x})^2$, variance $s^2 = \frac{\sum (x-\bar{x})^2}{N}$, standard deviation $s = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$
Correlation coefficient $r_{xy} = \frac{cov(x,y)}{s_x s_y}$ = $r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{¿¿¿}}$, where X and Y are individual measurements of two variates, X and Y.
Linear regression Slope $b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$, intercept $a = \bar{y} - b\bar{x}$, where *x* and *y* are individual measurements of an independent variable *x* and a dependent variable *y*.
t-test for comparison of means with unequal sample sizes *t*-statistic $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$, where common variance $s_p^2 = \frac{\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} + \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{m}}{n+m-2}$ and $\bar{x}_1 \wedge \bar{x}_2$ are means of two samples of size *n* and *m*.

Box 2. Analysis of variance table.

Source of variation	Corrected *SS*	*df*	MS	Variance ratio
Between (treatment)	SS_{bet}	$k-1$	$\frac{SS_{bet}}{k-1}$	$\frac{MS_{bet}}{MS_{wit}}=F_{(k-1,N-1)}$
Within (error)	SS_{wit}	$(N-1)-(k-1)$	$\frac{SS_{wit}}{(N-1)-(k-1)¿}¿$	
Total	SS_{tot}	$N-1$		