From:

Koch AL (1994). Growth measurement. *In:* Gerhardt P, Murray RGE, Wood WA & Krieg NR (eds) Methods for general and molecular bacteriology. American Society for Microbiology, Washington, D. C., pp 248-277

11.7. STATISTICS, CALCULATIONS, AND CURVE FITTING

Several aspects relevant to the measurement of bacterial growth are almost never included in courses in basic microbiology or statistics taken by undergraduate microbiology majors. For the former, this is because the needed material is a little beyond the scope of the courses, and for the latter, it is because the counts of colonies or particles is an all-or-none event and of lesser utility for most users of statistics. The following description is intended to make these mathematical tools readily available.

11.7.1. Population Distributions

The **binomial distribution** describes the chance of occurrence of two alternative events. For example, it provides the answer to the question, "What is the chance of having five boys in a family of nine, assuming that births of boys represent 0.56 of the total live births?" The numerical answer is $P_5 = 0.2600$. The relevant formula is as follows:

$$P_r = \frac{n!}{(n-r)!r!}p^r(1-p)^{n-r}$$

where p is the chance of a specified response on a single try, n is the total number of trials, and r is the number of specified responses and would vary from 0 to n. The symbol, !, means factorial; i.e., $n! = n(n-1)(n-2) \ldots 1$. The formula shows that in different families of nine children more families would have five boys than any other number. For example, $P_4 = 0.2044$, $P_5 = 0.2600$, and $P_6 = 0.2207$. The plot of P against r is an example of a distribution histogram. The binomial distribution provides a way to estimate the mean and standard deviation of p from experimental data. Assuming that there are data on only a single family with n children and that there are by chance r boys in that family, statistical theory shows that the best estimate of p is r/n. Numerically for the example where $r = 5$ and $n = 9$, p is $5/9 = 0.5556$ and its standard deviation is

$$\sqrt{p(1-p)/n} = \sqrt{(5/9)(4/9)/9} = 0.1656$$

This result is not very reliable because the coefficient of variation (CV) is nearly 30% (CV = 0.1656 × 100%/0.5556 = 29.81%). The precision in estimating p could be improved by looking for a family which had many more children and applying the same formula. It would be not only easier but better to pool the census data of a number of families. Suppose 50,000 boys are counted in 90,000 children in a large pool of families. Then the same formula yields $p = 0.5556$, but now CV will be 0.2981%. Not only is this more precise because large numbers are counted, but also many families, for genetic and sociological reasons, may have different values of p. *It may be desirable to estimate the value of p that applies generally to the entire population.*

The formula for the distribution is cumbersome to calculate when the numbers are large. An important contribution of Gauss was to rewrite the binomial distribution for this large-number case. Thus, the **Gaussian distribution** applies as a generalization of the binomial distribution for the case in which the numbers involved are so large that they can be treated as a continuous distribution instead of one with discrete variables. The variables of the Gaussian distribution replacing n and p are the population mean (m) and standard deviation (σ). The formula then becomes

$$P_x = \frac{1}{\sqrt{2\pi\sigma}}e^{-(x-m)^2/2\sigma^2}$$

where e is 2.71828.

In this formula the continuous quantity (x) replaces the positive integer variable (r) as the measurement of response. This distribution, like the binomial, can be mathematically manipulated so that one can go from data to estimations of these two parameters. The Gaussian distribution is also called the **normal distribution,** partly because it is symmetrical about the mean and partly because it is so frequently observed.

For data that follow a Gaussian distribution, the estimate of the mean is called m and is given by $m = \Sigma x/n$. The estimate of the standard deviation is called s and is given by:

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

The coefficient of variation, CV, is given by s/m.

The other limiting distribution of the binomial is the **Poisson distribution.** It applies for the case where n is very large and p is very small but the product np is finite. The best estimate of np is N, the observed number of total responses of a specified kind. This distribution would be useful if, for example, boys occurred very rarely (say, 1 in 10,000) but families were large (say, 100,000 children). Then, an average family would contain $N = 10$ boys and the standard deviation would be

$$n\sqrt{p(1-p)/n} = \sqrt{n(N/n)(1-N/n)} =$$

$$\sqrt{10{,}000\cdot(10/10{,}000)(1-10/10{,}000)} = 3.1621$$

A keen simplification, due to Poisson, was to assume $N/n << 1$. Then the formula for the binomial distribution simplifies to a one-parameter distribution

$$p_r = \frac{e^{-m} \cdot m^r}{m!}$$

The best estimate of the mean is $m = N$, and the best estimate of the standard deviation is $\sqrt{N}$. Note that this is not much different from the binomial distribution, since $\sqrt{10} = 3.1623$ is not much different from 3.1621. Note that again n and p are replaced by different symbols, in this case by a single one, m. The important point is that *the count of the number of discrete objects provides not only the best estimate of the mean value (i.e., $N = x = m$) but also an estimate of the precision of the estimate ($\sqrt{N} = s = \sigma$).*

Consequently, in the enumeration of objects it does not matter how they are subdivided. Two replicate plates with 200 colonies on each are no better or worse from the point of view of Poisson statistics than is one plate with 400 colonies. In both cases the standard deviation, s, of the measurement is, $\sqrt{400} = 20$ and the CV is $\sqrt{400}/400 = 5\%$. Therefore, to get the best estimate from a group of plates from the same or different dilutions of the same sample, simply add the total counts on all the plates and divide by the total volume of the original solution. The standard deviation is the square root of the total count divided by the plated volume of solution.

As an example, imagine that duplicate plates were made at dilutions of both 10^{-5} and 10^{-6} with counts of 534 and 580 and of 32 and 60, respectively. The total count is 1,206. If 0.1 ml of these dilutions were plated, 2.2×10^{-6} ml would be the total volume of original culture used to make the four plates. Therefore, the best estimate of the concentration is $1{,}206/2.2 \times 10^{-6} = 5.46 \times 10^8$/ml. The standard deviation is $\sqrt{1{,}206}/2.2 \times 10^{-6} = 0.16 \times 10^8$/ml.

Justification for treating the results at the two different dilutions separately and not pooling them, as done above, depends on other kinds of errors being larger. Then, by comparing the results at different levels of dilution, some estimate is obtained of the variability due to the additional pipetting operation and to other sources of error that are not included in the calculation of the Poisson sampling error. Although this variability is of interest, sources of error could be more directly measured by carrying out independent dilutions from the same cell suspension. As an example, imagine that 0.1-ml aliquots were plated on single plates from each of 12 independent 10^{-5} fold dilutions and that the following set of colony counts were obtained: 534, 580, 760, 643, 565, 498, 573, 476, 555, 634, 514, 693. The sum is 7,026, and the Poisson standard deviation is $\sqrt{7{,}026} = 83.8$. The mean of the numbers is 585.5, and the standard deviation by the Gaussian formula is 81.2. Calculation of the bacterial count of the original suspension together with the two different estimates of error yields count $= 7{,}026/1.2 \times 10^{-5}$ ml $= 5.85 \times 10^8$/ml; Gaussian error $= \pm 81.2/0.1 \times 10^{-5} = \pm 0.81 \times 10^8$/ml; and Poisson error $= \pm 83.8/1.2 \times 10^{-5} = \pm 0.07 \times 10^8$/ml. The comparison of these two estimates of error suggests that considerable error is due to sources other than random sampling. An attempt should be made to find and reduce these sources of error. Until that is done, it is necessary to make many independent dilutions and use Gaussian statistics. *Until the source of error is found, the Poisson error is irrelevant.*

This same point can be made in another way from this example. Imagine that only one plate had been made, say the first one, in which case only the Poisson error would be available for consideration. The count then would be $5.34 \times 10^8 \pm 0.23 \times 10^8$, and the real error would be underestimated fourfold. It is therefore cautioned not to rely on Poisson statistics until their use has been justified for the conditions actually in use. Instead, make a comparison on at least several occasions with a Gaussian statistic measurement of error as indicated above.

11.7.2. Statistical Tests

Much of the statistics taught in elementary courses is concerned with whether a body of data is consistent with a hypothesis. Usually the **probability** (P) that the observed deviations from the hypothesis could occur by chance is computed. If P is small, but not very small, the hypothesis could still be false although improbable. If P is very small, the hypothesis can, with some confidence, be said to be true. These statistical tests are generally made on the assumption that the data follow a Gaussian distribution. In many cases in bacteriology, this assumption should be questioned and other appropriate statistical tools should be used.

The standard deviation has been defined above. This is frequently confused with another term, the **standard error,** also called **standard deviation of the mean.** The standard deviation measures the deviation of an individual measurement from the mean of many measurements. The standard error measures the mean of all the data observed from the mean of a hypothetical data base containing an infinite number of observations and is a measure of how close the average is to the "true" mean value.

The only statistical test mentioned in the text below is Student's ***t* test**. This applies to the difference in the means of two groups. The difference is divided by the standard error of the combined data. Thus the t value measures the difference in the means by using the standard error as its unit of measurement. This ratio is compared with values given in tables. Use of the tables generates P values but requires a knowledge of the number of measurements and whether potential deviations can occur on both sides or only one side of the mean. The bigger that t is, the smaller P becomes; if P is not very small, the hypothesis that the two populations were identical may not be rejected.

In recent years, the **analysis of variance** (ANOVA), which is a subbranch of statistics, has been elaborated so that it now can be applied to many problems and replaces many of the more specialized techniques used previously. The availability of packaged programs for various computers means that it requires work, but much less work than previously, to learn to use and apply statistical methods when they are appropriate in bacteriology.

11.7.3. Error Propagation

The accuracy of an estimate depends on the accuracy of its component measurements. The Poisson error of a colony count and the error of the dilution procedure both contribute to the error in the estimated concentration of organisms of the original undiluted suspension. Additional errors can only further blur the results or make them less precise. Even though errors in one part of the estimate may compensate for errors in another part, on the average random errors will make them larger. When errors in one measurement are independent of (uncorrelated with) errors in another measurement, the overall error can be calculated by two rules for **propagation of errors,** as follows.

1. If two quantities (x and y) are to be added or subtracted, the standard deviation (s) of the combined quantities is

$$s_{x+y} = s_{x-y} = \sqrt{s_x^2 + s_y^2}$$

2. If two quantities are to be multiplied or divided, the coefficient of variation (CV) of the combined quantities is

$$\mathrm{CV}_{xy} = \mathrm{CV}_{x/y} = \sqrt{\mathrm{CV}_x^2 + \mathrm{CV}_y^2}$$

As an example, apply the second rule to estimate the overall error in a single plate count containing colonies from the series of 10^5-fold dilutions. Assume that the dilutions were performed in five steps of 10-fold each and that the pipetting error of a single 10-fold dilution has a CV of 0.02. Then the overall error of the five dilution steps is $\sqrt{5} \times 0.02$. This result is obtained by the repeated use of the second rule. It then must be combined with the Poisson error. Since the best estimate of the Poisson error CV is $1/\sqrt{585.5}$, then the overall CV is as follows:

$$\mathrm{CV} = \sqrt{1/585.5 + 5(0.02^2)}$$

$$= \sqrt{0.001709 + 0.00200} = 6.1\%$$

This 6.1% error is composed of a Poisson counting error of 4.1% = $100/\sqrt{585.5}$ and an error due to the cumulative pipetting errors of 2% × $\sqrt{5}$ = 4.47%. The rule to combine them gives a value smaller than their sum (4.1 + 4.47 = 8.57%) but larger than the largest contributor to the error.

Two important experimental considerations derive from this example. First, *there is no reason for increasing the accuracy of one part of an experiment unless other sources of error comparable to it are also decreased or eliminated.* Second, *if an operation is to be done many times, it is worthwhile to devise a way to do it accurately and then obviate the need to carry out elaborate statistical calculations.* In the previous section, the pipetting error was neglected because it was assumed that pipetting can be and was done accurately. This is a reasonable thing to do if the CV of this error is smaller than the Poisson counting error. As an example, imagine that each pipetting operation had been carried out with an accuracy of 1% instead of 2%. Then the overall pipetting error would have been 1% × $\sqrt{5}$ = 2.23% and the overall total CV consequently would have been $\sqrt{4.1\%^2 + 2.23\%^2} = 4.7\%$, only a little bit larger than the Poisson counting error by itself.

Similar logic follows for cases in which blank values and background values are to be subtracted (in which case the first rule for propagation of errors applies) or in which the instrument calibration factors are used to multiply observed values (and the second rule applies). In the measurement of controls used repeatedly in the calculation of data, errors should be reduced by repetitions or by more accurate measurement than for individual experimental values so that the control factors do not contribute significantly to the overall error of measurement.

11.7.4. Ratio Accuracy

There is a very powerful and general statistical method applicable to diverse experimental situations varying between large natural ecosystems at one extreme and a drop of culture on an electron microscope grid at the other. This method is to add a known number of reference particles, which may be bacteria, ferritin particles, polystyrene spheres, abortively transduced bacteria, plasmids, viruses, etc. After mixing takes place, samples are taken and the ratio of the number of objects of interest to the number of reference particles is determined by appropriate means. The method can be illustrated for microscopic smears containing a class of recognizable organism of unknown number and polystyrene beads of known concentration. The smear must be prepared from a known volume of cellular suspension and a known volume of suspension of beads of known concentration. The concentration of the beads in the final suspension is multiplied by the ratio of the counts of the unknown cells relative to those of the reference beads to calculate the concentration of unknown cells. The second rule for the propagation of error applies in this case. If the concentration of the reference particles is known without error in the original stock solution, the coefficient of variation of the unknown particles is given by

$$\mathrm{CV} = \sqrt{\frac{1}{N_u} + \frac{1}{N_r}}$$

where N_u and N_r are the counts of the unknown and reference, respectively. To minimize the number of total counts following the first argument in the previous action, N_u should be about equal to N_r. Then the CV will be about $\sqrt{2}$ = 1.4-fold larger than if a very large number of reference cells (or unknown cells) had been counted.

11.7.5. Coïncidence Correction

Coïncidence corrections must be applied when too many colonies are on a plate or too many cells are on a square of a counting chamber or if too many radioactive decays are recorded by a radioactivity counter in a unit of time. For the case of a colony count, assume that, if two cells initially are closer together than a distance r, they will be counted as a single colony. Let N_t be the true count and N_a be the actual count, and assume that the radius of the petri dish is R. Consider a single cell; the chance that another cell on the plate is within a distance r is $N_t r^2/R^2$; thus, the count is decreased by $N_t r^2/R^2$. The number of colonies not counted will be $N_a N_t r^2/R^2$. Therefore,

$$N_t = N_a + N_a N_t r^2/R^2 = N_a(1 + cN_t)$$

where $c = r^2/R^2$. It is usually convenient to substitute N_a for N_t on the far-right-hand side when the correction is small. From this formula it is clear why a fourfold reduction in colony size reduces the coincidence correction at a given count by 16-fold. This is the basis of the use of layered plates (section 11.3.3), which makes the colonies smaller and thus reduces coincidence.

11.7.6. Exponential-Growth Calculations

Under constant conditions after a long enough time when cell-cell interaction is small, growth measured in any manner is expected to proceed according to $X = X_0 e^{\mu t}$. This can be written in any of the following equivalent ways:

$$\ln X = \ln X_0 + \mu t$$

$$\log X = \log X_0 + \mu t/2.303$$

$$X = X_0 2^{t/T_2}$$

In the last equation, T_2 is the **doubling time** and can be calculated from the following:

$$T_2 = (\ln 2)/\mu = 0.6931/\mu$$

Many symbols other than μ have been used for the specific growth rate including a, k, and λ. Knowing these other symbols is important because they are used without definition in many papers. Confusion arises with μ, which designates the specific growth rate in the literature on continuous culture (Chapter 10.2) and in microbial ecology. This is the usage employed here. Unfortunately, μ is also used to symbolize the number of doublings per hour in the literature on cell physiology. The latter usage differs from the former usage by a factor of ln 2 = 0.6931. Although any time unit could be used, reciprocal hours appears to be nearly standard. The doubling time (T_2) is reported in the literature in either minutes or hours.

11.7.7. Plotting and Fitting Exponential Data

There are several alternative ways to fit data to the exponential-growth model that are equally valid but differ in their precision and in the additional information yielded to the experimenter. The most simple conceptually is to look up the natural logarithms of the concentration of cells (or the dry weight of biomass, or other measurement of an extensive property of the bacteria) and plot them on ordinary arithmetic graph paper against the time that the measurements were made. Then draw a straight line through the datum points as close to the points as possible, and determine its slope by the rise-over-fall method. If the time scale is in hours, then the slope is in units of reciprocal hours. At this point, ask two questions: How appropriate is a straight line to the data? Is the line drawn a good summary of the data?

Common (base 10) logarithms may be used, in which case the slope must be multiplied by 2.303 = ln 10 to obtain μ. Base 2 logarithms also may be used, in which case multiply by 0.6931 = ln 2. There is an advantage in using base 2 logarithms: the reciprocal of the slope gives the doubling time directly. Whatever the base of logarithm used, plot both the characteristic and mantissa numbers on the arithmetic scale of the graph paper. Do not make the all-too-frequent blunder of mixing them (e.g., $\log_{10}$ of 4×10^8 cells = $8.6 \neq 8.4$)!

It usually is more convenient to use semilogarithmic graph paper. Paper that has six divisions between darker lines on the arithmetic scale and two cycles on the logarithmic scale is recommended. Define the abscissa (x axis) according to hours or days on the major lines so that an even number of minutes or hours is represented by the minor lines. Mark each of the three unit labels with the appropriate powers of 10 on the ordinate (y axis). The printed scales may be multiplied by a constant, but it is invalid to add or subtract a constant from the logarithmic scale. Find the point corresponding to the amount of biomass, cell numbers, or other measure of extent of growth on the y axis, and mark the point exactly. It is useful to make a small point mark for exactness and surround it with a larger circle for better visualization. It also is useful, when light-scattering methods are used, to plot each point as soon as it is obtained. This frequently shows when errors have been made, whether they be biological, instrumental, arithmetic, or human. If the error is detected immediately, one has the opportunity to restart a culture, remeasure the culture, or replot a point. Once all the data have been plotted, draw the best straight line to fit the points.

The mathematical procedure that does this best is called the **least-squares** fit, which minimizes the square of the vertical distance of all points to the proposed line. This can be approximated visually by mentally noting the distance from the few points that are farthest from the position of a proposed straight line by moving a transparent ruler. By readjusting the ruler and remembering that the actual distances, even though the graph paper is semilogarithmic, are to be squared, one can do a quite accurate job of drawing very nearly the line which would be generated by the mathematical least-squares procedure.

Figure 7 shows an actual growth curve with an "eyeballed line" which is essentially the same as the computer-fitted line. This example is drawn from a carefully executed experiment with accurate data over a period permitting a 30-fold increase in cell mass. Note that the datum points fit close to the line. This implies that if overt or gross errors were made during the experiment, they were discovered. This example was chosen to show how sensitive growth rate measurements are to the accuracy of the data and to show the way the line is drawn. The drawn line is almost as good a representation of the data as (and is indistinguishable from) the computed line over the range from 10 to 20 μg (dry weight)/ml but corresponds to a 2.7% difference in μ or T_2.

Note that there are many kinds of semilogarithmic graph paper classified by how many cycles (powers of 10) they span. The fewer the number of cycles, the more spread out the points are and the easier it is to define the line accurately. Note also that there is only one type of