Counting bacteria: confidence intervals from one measurement

Valentin - http://zulko.wordpress.com/

November 14, 2012

Here are two little biologically-inspired statistics problem, with very simple solutions. They show that, when you know how your data is collected, one observation may be enough to give a confidence interval.

1 First Problem

Problem From a bacterial solution, I sampled a small volume v, in which I found no bacteria. Give an upper bound (with confidence 95%) for the concentration c of bacteria in the solution.

1.1 Solution

We will show that $c < 3/v$ with (slightly more than) 95% confidence. Let us call V the number Suppose that there are N bacteria in total in your solution. Whitout further indications, we can suppose that the solution is homogenous (if not, shake !). So the probability of each of the N bacteria to be in your small sample is (v/V). The probability that none of these bacteria will be in the sample (what was observed) is then $(1 - v/V)^N$. Now, the 95%-confidence upper bound for N is the largest possible N that keeps the probability of the observed event above 5%. Said otherwise it is the largest N such that

$$(1 - v/V)^N > 0.05$$

Applying the log function to both sides we get

$$N \log(1 - v/V) > \log(0.05)$$

and further (beware that $\log(1 - v/V) < 0$)

$$N < \frac{\log(0.05)}{\log(1 - v/V)}$$

Here we have found an exact upper bound, which is nice, but it can be simplified in two ways. First, if v/V is small, then

$$\log(1 - v/V) = -v/V - (v/V)^2 - (v/V)^3 - ... \lessapprox -v/V$$

The second trick is much less known, yet very helpful:

$$\log(0.05) = -2.9957... \gtrapprox -3$$

This leads to the very tight upper bound:

$$N < \frac{\log(0.05)}{\log(1 - v/V)} < \frac{3}{v/V} = \frac{3V}{v}$$

Hence the result : $c = N/V < 3/v$

2 Second Problem

Problem From a bacterial solution, I sampled a small volume v, in which I found n bacteria ($n > 20$). Give an upper bound (with confidence 95%) for the concentration of bacteria in the solution.

2.1 Solution

The fact that we found n bacteria in a volume v suggests that in the solution has a bacterial concentration of n/v. We will show that

$$[\frac{n \pm 2\sqrt{n}}{v}]$$

is a good (approximated) 95% confidence for N. As explained in the previous section, the probability of each of the N bacteria of the solution to be in the sample is v/V, where V is the volume of the solution. So the number of bacteria in the sample will follow a binomial law with parameters $(N, p = v/V)$. Since Np is big enough[1] one can approximate this law with a normal law of mean Np and standard deviation $\sqrt{p(1-p)N}$. If we call σ the 97.5%-quantile of the standard normal distribution, then the number of bacteria in a sample of volume v has a 95% probability of falling in $[Np \pm \sigma\sqrt{p(1-p)N}]$. The 95% confidence interval for N is the interval containing all the values of N such that the observed value n belongs to this interval, i.e. (cf. figure)

$$Np - \underbrace{\sigma\sqrt{p(1-p)}}_{b}\sqrt{N} \leq n \leq Np + \sigma\sqrt{p(1-p)}\sqrt{N}$$

[1] Approximating a binomial law with a normal law requires $Np > 10$. If Np was smaller than 10, then the standard deviation of n would have been at most $\sqrt{10}$ and the probability of observing more than 20 bacteria would have been very low (said otherwise, we can rule out $Np \leq 10$)

If we denote $x = \sqrt{N}$, then x must verify the following inequalities

$$\begin{cases} 0 < x \\ 0 > px^2 - bx - n \\ 0 < px^2 + bx - n \end{cases} \quad (1)$$

The solution of which is

$$\frac{\sqrt{\Delta} - b}{2p} < x < \frac{\sqrt{\Delta} + b}{2p}$$

where $\Delta = b^2 + 4np$. From this we deduce

$$\left(\frac{\sqrt{\Delta} - b}{2p}\right)^2 < N = x^2 < \left(\frac{\sqrt{\Delta} + b}{2p}\right)^2$$

which leads to

$$\frac{2b^2 + 4np - 2b\sqrt{\Delta}}{4p^2} < N < \frac{2b^2 + 4np + 2b\sqrt{\Delta}}{4p^2}$$

In terms of the original variables:

$$N \in \left[\frac{v}{V}\left(n + \frac{1}{2}\sigma\sqrt{1 - \frac{v}{V}} \pm \frac{1}{2}\sigma\sqrt{1 - \frac{v}{V}}\sqrt{\sigma^2(1 - \frac{v}{V}) + 4n}\right)\right]$$

We found an interval that answers the problem, but its expression is a little heavy. To simplify it we will use the approximations $\sigma = 1.96 \simeq 2$ and $1 - p \simeq 1$ (since $p = v/V$ and v is supposed small compared to V). This leads to

$$[(V/v)(n + 1 \pm 2\sqrt{n + 1})]$$

Which can be further approximated to get the announced result.

Remark The idea here was to give a solution as general as possible, then approximate it to obtain a handy formula. Note that the result could have been obtained in a simpler (but not as rigorous) way by considering that since the number n follows a binomial law of mean Np and variance $Np(1 - p) \simeq Np$, and that n is a good approximator for nP, we can say that n was obtained using an estimator of mean n and standard deviation $\sqrt{n}$ and hence the *real* n should be actually somewhere between n and two standard deviations. This is a short way of coming to the result, but using a wrong reasonning. In particular, it doesn't acknowledge the fact that the variance of the estimator, in our problem, is not independent from its mean !

3 Cave at

All these considerations rely on the hypothesis that the solution is homogenous, which may not be the case (because bacteria aggregate, make biofilms, and so on). The best way to check the validity of this hypothesis is to make replicates. If you find that the variance $\sigma_k$of your k replicates is larger than the square-root of their mean, $\sqrt{n_k}$, then it may mean that this hypothesis is not valid (shake better !).