

STANDARD ERRORS

Don't get t out of proportion!

Gerald Goodall

Brunel University, Uxbridge, England.

WHEN n is large, confidence intervals for the p parameter of a binomial distribution may be based on the quantity given in the usual notation as

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \quad (1)$$

whose distribution is (approximately) $N(0,1)$. Thus a 95% two-sided confidence interval is given (approximately) by

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad (2)$$

and similarly for one-sided intervals and different confidence coefficients.

A mistake that is quite commonly found when marking students' coursework and when marking public examination scripts is to suppose that the quantity (1) has a t distribution, with an assorted number of degrees of freedom, so that the confidence interval is given by an expression similar to (2) but with a value from t instead of $N(0,1)$.

It is not difficult to see how this mistake arises.

When finding a confidence interval for the mean μ of a Normal distribution we use

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{when the variance } \sigma^2 \text{ is known, this}$$

quantity having a $N(0, 1)$ distribution. When σ^2 is unknown, we estimate it using the usual "sample variance" S^2 (divisor $n-1$) and then base our interval on

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

which has a t distribution, with $n-1$ degrees of freedom. In the binomial proportion case, the expression with which we work also contains an unknown parameter (p itself) that has to be estimated. It is, on the face of it, a reasonable analogy with the case of the interval for p to suppose that a t

distribution, rather than $N(0, 1)$, should again be used.

The purpose of this note is to try to explain why this is not so.

To see this, we need first to look at the μ case more thoroughly to try to find out *why* the t distribution is appropriate here if the variance is estimated. The result is a standard piece of statistical theory at a more advanced level than would be taught in school. It would not be possible to *prove* the result in school, because it depends on quite advanced university-level mathematics. Nevertheless it is a result that might well be *stated* and *explained* which would hopefully go a long way to overcoming the misunderstanding about the p case.

The result has two parts of which the second is in a sense the easier to state and explain. This is the *formal definition* of the t random variable with m degrees of freedom as the quotient of a $N(0, 1)$ variable and the square root of a χ -squared variable with m degrees of freedom itself divided by m , where the $N(0, 1)$ and the χ -squared are independent. Rather a mouthful in words, but easy to turn into symbols:

$$t_m = \frac{N(0,1)}{\sqrt{\frac{\chi^2}{m}}}$$

Using this definition, the density function of the t random variable can be found from those of the Normal and χ -squared variables using two-variable transformation techniques - a quite advanced piece of work.

We still need to see why this definition is helpful and relevant for the μ case. The first part of our required result now comes into play - that the distribution of the sample variance for observations from a Normal distribution is given by χ -squared in the following way:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

$$\text{where } S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

Proof of this is arguably yet more advanced again, depending on theory of quadratic forms in random variables. But the *use* of this result is actually fairly common in A-level syllabuses, at least at Further Mathematics level, where the χ -squared test for the variance of a Normal distribution is often a syllabus item. If the result has been stated for that purpose, it might as well be used for the present purpose too!

Putting these pieces of work together, we now argue as follows. Using the familiar result for the μ case when the variance is known, we have that

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

and we now also have that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

so that

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

This, then, *derives* the *t*-based confidence interval for μ when the variance is unknown. It is no longer a piece of ‘magic’ to the effect that “whenever there are estimated parameters, the *t* distribution has to be used”. On the contrary, the proper theoretical derivation is now evident. Agreed, it is still partly hidden behind advanced results that cannot yet be proved, but nevertheless the *idea* of how the theory goes is clear. It ought to be able to be understood by sixth-formers seriously interested in statistics, and even getting this far is much to be preferred to leaving it as ‘magic’.

For we can now see immediately that the *t* distribution is simply *irrelevant* in the case of the *p* confidence interval. The quantity (1) is not of anything like the required form to generate the definition of the *t* distribution. So we use (1) to obtain the *p* interval *but still using* $N(0,1)$ *as the (approximate) underlying distribution*.

