HYPOTHESIS TESTS, CONFIDENCE INTERVALS AND COMMON SENSE

Timothy E. O'Brien
Loyola University Chicago, USA
& Katholieke Universiteit Leuven, Belgium

Applied researchers are often interested in obtaining confidence intervals for key nonlinear model parameters so as to answer important research questions, and the usual "plus and minus 2 SE's" confidence interval leads easily into the usual Wald hypothesis test covered in most introductory statistics courses. However, since information about a specific parameter is often asymmetric, a skewed confidence interval is often more appropriate and reasonable in practice. This leads to the use of likelihood-based tests, typically introduced in intermediate undergraduate and basic graduate course. This paper argues that the superiority (in terms of for example increased power) of likelihood-based and score hypothesis tests over the Wald test is most easily conveyed and appreciated by first providing a reasonable motivation (as well as examples) using confidence intervals, and then exploiting the confidence interval-hypothesis test equivalence.

INTRODUCTION

One of the more interesting applications of statistics encountered by students in a basic statistics or biostatistics course is that of tests of significance, either indirectly by using confidence intervals or directly by using hypothesis tests. Indeed, instructors often draw connections between confidence intervals and hypothesis tests – for example when testing whether average grade point averages differ for independent samples of male and female students, and under the usual assumptions (i.e., Gaussian distributions and equal variances); see for example Samuels and Witmer (1999). That is, in simple or large-sample situations, the usual Wald confidence interval ("estimate +/- 2SE's") is sufficient in the sense that the actual coverage and the nominal coverage (e.g., 95%) usually coincide. Yet, even in introductory courses, Wald intervals are signaled as potentially problematic; e.g., confusing adjustments are recommended to Wald intervals for a single binomial parameter in Samuels and Witmer, and students are then led to employ methods that are confusing and often resort to memorization. Similarly, the methods developed for significance testing in follow-up and advanced course are often equally obscure.

This paper underscores that such situations provide a wonderful opportunity for instructors to permit interested students to develop a deeper understanding of important issues in hypothesis testing, perhaps in the form of a course project. Significance testing approaches are discussed in the following two sections for students enrolled in a first course or in a follow-up course followed by various practical illustrations. As interested students may wonder how Wald intervals can be adjusted to bring them more in line with exact intervals, we also address marginal curvature and design of experiments, topics that provide a clear motivation for interested students to take additional follow-up courses.

STRUCTURING TOPICS IN A FIRST STATISTICS COURSE

Although Chance and Rothman (2001) provide arguments for the teaching of hypothesis testing before confidence intervals for one- and two-sample problems, commonly used textbooks such as Moore and McCabe (1989) and Samuels and Witmer (1999) reverse this order. Indeed, although many authors feel that hypothesis testing follows logically just after a discussion of sampling distributions, most authors and instructors usually find it prudent to alternate between testing and intervals as focus shifts from comparing two population means to comparing two population proportions, and so on. Students in a first statistics course can easily become perplexed, however, as alternative methodologies are proposed for interval estimates even for a single binomial proportion or for a correlation coefficient; note, for example, the Wilson technique discussed in Samuels and Witmer, q-intervals discussed in Santner (1998), and the resampling techniques proposed in Johnson (2001) and Ricketts and Berry (1994). Clearly, in these situations, students naturally wonder if the interval-significance test analogy still holds.

Moreover, an important matter such as the inappropriateness of Wald intervals in certain situations provides an excellent topic on which students can focus course projects (commonly

required of statistics students – for example those enrolled in the author's beginning biostatistics course) since it allows students to obtain a heuristic understanding of likelihood-based intervals, curvature and adjustments to the Wald intervals to bring them more in line with the likelihood intervals. Since the poor coverage of the Wald interval is best introduced by means of concrete illustrations, we provide several examples on which students could focus this project. First, we discuss similar problems encountered in follow-up courses.

STRUCTURING TOPICS IN COURSES BEYOND THE FIRST

Basic courses in regression, categorical data analysis and intermediate (biostatistical) methods often focus on topics such as nonlinear, logistic and log-linear regression, relative risk, odds ratios, amongst others – topics which typically are not be covered in basic courses of statistics. Tests of significance for these situations and models are usually first encountered via hypothesis testing (see, e.g., Agresti, 1996; Christensen, 1997; Gallant; 1987), and students are often very confused as to why likelihood-based and score tests are preferred to Wald tests. Although an instructor may argue the superiority of the former tests in terms of statistical power, this latter concept is somewhat advanced and abstract for those students interested in applications and less interested in statistical theory. Equally confusing to intermediate-level students are the transformations involved in obtaining confidence intervals for correlation coefficients, odds ratios and the relative risk (see, e.g., Samuels & Witmer, 1999; Agresti, 1996).

It is argued here that the preference for likelihood-based testing procedures over Wald-based procedures is best conveyed by first illustrating the more reasonable coverage properties of likelihood-based confidence intervals as compared with Wald intervals. It can also be pointed out to students that omitting score methods from the discussion is reasonable in light of the words of caution provided in Bates and Watts (1988), who strongly state, "we do not recommend the approach"; the undesirable properties of the score, or exact-sampling, method result from the fact that both the corresponding numerator and denominator vary with the parameters. Further, transformations to Wald intervals for binomial proportions, correlation coefficients, odds ratios and the like can then be illustrated as providing intervals which are more in line with the more-preferred likelihood-based intervals.

SOME EXAMPLES

A first example which illustrates the superiority of likelihood methods, provided in Zar (1999), involves $y = 4$ successes in $n = 500$ (independent) trials in estimating the underlying binomial proportion. For this example the likelihood-based confidence interval (LBCI) for p extends from 0.006 to 0.046, whereas the Wald interval (WCI) runs from 0.001 to 0.039. This represents an overlap of less than 73%; *overlap* here is defined as the intersection of the intervals divided by the union, as used in van Ewijk and Hoekstra (1994). Zar also discusses a transformed interval provided by Bliss and Brownlee, and which provides a better approximation to the LBCI. Although this is a rather simple example, it serves to illustrate the divergence between these intervals and the need for transformations in certain situations, especially when it would be impractical (or impossible) to calculate the likelihood interval.

The quadratic regression model, $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$, discussed in Cook and Weisberg (1990), provides a second illustration. If θ ($= - \beta_1 / 2 \beta_2$) represents the x-value of the maximum and is a parameter of interest, this model becomes a nonlinear one (in the parameters), as it is now written, $E(Y) = \beta_0 - 2\theta\beta_2 x + \beta_2 x^2$,. Using a data-set similar to that considered by Cook and Weisberg yields a point estimate of θ of 0.186, a WCI extending from –0.048 to 0.420 and a LBCI from -0.513 to 0.333, which represents a marked difference (the overlap of these intervals is just 40.8%). The reasonableness of the latter interval over the former one is best seen by noting that since the information provided by the data for the parameter of interest is very asymmetric (with much more information on the right-hand side than on the left), so should the interval be asymmetric and thus shifted to the left – as is the case for the likelihood interval. Such an example enables students to easily accept the superior performance of likelihood-based testing procedures over Wald procedures. It is therefore extremely helpful to students to first address confidence intervals before hypothesis tests in courses beyond the most basic. Further, students

can easily obtain LBCI's using SAS, and programs provided by the author can be downloaded from the following website:
www.math.luc.edu/~tobrien/courses/glm/SAS-Programs/PLCI-Lansky2-model.html

Finally, instructors of intermediate-level biostatistics courses may want to provide another example involving the relative risk (RR) and the transformed interval (ACI) provided in Agresti (1996). For 5 successes in a group of size 75 and 20 successes in a (independent) group of size 25, RR is estimated 12.0, and the LBCI is (5.60,33.23) whereas the ACI extends from 5.03 to 28.62. This example points out that even transformations may yield only an 82% overlap, and since LBCI have better coverage properties, hypothesis tests based on likelihood methods are often much preferred.

With these examples as an introduction and motivation, students will easily appreciate the strong preference for the likelihood ratio significance test to the Wald test.

A FORAY INTO CURVATURE AND DIFFERENTIAL GEOMETRY

Interested students will often wonder why Wald and likelihood intervals may diverge, and instructors can focus a short discussion on the heuristics of curvature and differential geometry – especially to the more mathematically sophisticated students – pointing out that Wald methods (as well as most commonly-used computing algorithms) are based on tangent plane approximations to underlying curved expectation surface. More advanced students may also wish to read through some of the details provided in Chapter 7 of Bates and Watts (1988) and Chapters 4 and 5 of Seber and Wild (1989). Finally, although many of the technical details provided in Clarke (1987) might only be understood by advanced students, all students are often interested to learn that Clarke's paper provides a means of adjusting the endpoints of the Wald interval to bring them more in line with the corresponding likelihood interval. SAS programs, available from the author, have been used by students to obtain these so-called marginal-curvature-adjusted intervals (MCCI). For example, for the nonlinear quadratic regression model discussed in the previous section, the MCCI is (-0.300,0.374), and thus does a much better job of approximating the LBCI (the overlap of these intervals exceeds 70%).

COMMENTS ON DESIGN OF EXPERIMENTS

When discussing the standard error for the slope parameter in a two-parameter linear model, instructors often point out that this standard error is reduced if the x values are more spread out since it involves the sum of the squared deviations (in the x's) in the denominator. A natural extension of this is to point out that one would want to reduce the standard error of each of the model parameters, which leads one to provide parameter confidence ellipses (instead of intervals) for various designs, and often leads instructors into comment on favoring one design over another, perhaps even mentioning the D-optimality design criterion. This latter criterion is then discussed at greater length in the corresponding course on experimental design; see, e.g., Chapters 13 and 14 of Box and Draper (1987) as well as Neter, Kutner, Nachtsheim, and Wasserman (1996).

For the examples provided above, however, since Wald methods proved insufficient, students can easily recognize the limitations of first-order design criteria (such as D-optimality), and this often leads to a brief discussion of quadratic design criteria, or criteria which attempt to simultaneously reduce the length of confidence intervals and reduce curvature; see Seber and Wild (1989:260).

DISCUSSION

It has been argued that the sequencing of topics of confidence intervals followed by hypothesis testing coupled with simple examples similar to those given above provide students with an easier understanding of the preference of likelihood methods over Wald methods. Further, simple examples such as those considered here also provide a wonderful opportunity to motivate a discussion of alternative (to the Wald) hypothesis tests and confidence intervals and statistical power (i.e., topics from statistical theory), of methodological tools such as logistic regression or Gaussian nonlinear modelling, of the basics of differential geometry and curvature, of statistical computing and convergence algorithms, and of the choice of the experimental design. Therefore,

these and other examples serve to help intermediate-level students and practitioners draw connections between otherwise disjoint topics and courses, and provide a wonderful conclusion to a basic statistics course as well as motivation to students to continue their studies into these fields.

ACKNOWLEDGEMENTS

The author thanks Martin Buntinas of Loyola University Chicago (USA) and Alan McLean of Monash University (Australia) for discussions and comments that resulted in improvements to this paper. This paper was completed while the author was Visiting Professor of Statistics at Katholieke Universiteit Leuven (Belgium), and the author is grateful for financial support from KUL.

REFERENCES

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Bates, D.M., & Watts, D.G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley.

Box, G.E.P., & Draper, N.R. (1987). *Empirical model building and response surfaces*. New York: Wiley.

Boyle, C.R. (1999). A problem-based learning approach to teaching biostatistics. *Journal of Statistics Education*, *7*, 1-19.

Chan, W-S. (1995). On large-sample tests concerning proportions. *Teaching Statistics*, *17*, 16.

Chance, B.L., & Rossman, A.J. (2001). Sequencing topics in introductory statistics: a debate on what to teach when. *The American Statistician*, *55*, 140-144.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. New York: Springer-Verlag.

Clarke, G.P.Y. (1987). Marginal curvatures and their usefulness in the analysis of nonlinear regression models. *Journal of the American Statistical Association*, *82*, 844-50.

Cook, R.D., & Weisberg, S. (1990). Confidence curves in nonlinear regression. *Journal of the American Statistical Association*, *85*, 544-51.

Gallant, A.R. (1987). *Nonlinear statistical models*. New York: Wiley.

Johnson, R. (1997). Earth's surface water percentage? *Teaching Statistics*, *19*, 66-8.

Johnson, R.W. (2001). An introduction to the bootstrap. *Teaching Statistics*, *23*, 49-54.

Moore, D.S., & McCabe, G.P. (1989). *Introduction to the practice of statistics*. New York: W.H.Freeman & Company.

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). *Applied linear statistical models*. New York: McGraw-Hill.

Ricketts, C., & Berry, J. (1994). Teaching statistics through resampling. *Teaching Statistics*, *16*, 41-4.

Samuels, M.L. & Witmer, J.A. (1999). *Statistics for the Life Sciences*. New Jersey: Prentice-Hall.

Santner, T.J. (1998). Teaching large-sample binomial confidence intervals. *Teaching Statistics*, *20*, 20-3.

van Ewijk, P.H. & Hoekstra, J.A. (1994). Curvature measures and confidence intervals for the linear logistic model. *Applied Statistics*, *43*, 477-87.

Zar, J.H. (1999). *Biostatistical Analysis*: New Jersey: Prentice-Hall.