CONFIDENCE CURVES: AN OMNIBUS TECHNIQUE FOR ESTIMATION AND TESTING STATISTICAL HYPOTHESES[1]

Allan Birnbaum
New York University

A standard practice of physical scientists is to report estimates ("measurements") accompanied by their standard errors (or alternatively "average errors" or "probable errors"). Such reports are interpreted flexibly as appropriate in various contexts of application. With the usual normality assumption, such reports may be read as representing confidence intervals or limits at the various confidence levels, and this omnibus character largely accounts for the convenience and flexibility of such reports and interpretations. For estimators not normally distributed, a formal analogue of such reports is provided by *confidence curves*, which are estimates of an omnibus form incorporating confidence intervals and limits at various levels. The definition and computation of such estimates, and their graphical representation and interpretation, are discussed and illustrated by an example.

1. INTRODUCTION AND SUMMARY

This note describes estimates of an omnibus form, called a *confidence curve*, which incorporates confidence limits and intervals at various levels and a median-unbiased point estimate, together with critical levels of tests of hypotheses of interest, and representations of the power of such tests.

Such estimates can be represented conveniently and interpreted flexibly, as seems appropriate for typical purposes of informative inference, that is, purposes of representing suitably the general significance of observations as evidence relevant to the determination of values of unknown parameters. The use of such estimates avoids the need for adoption of a particular confidence level, which is overly schematic for such purposes. Much of current practice in applied statistics involves flexible use and interpretation of confidence intervals and confidence coefficients; confidence curve estimates provide an explicit unified form representing such practice. Estimates of omnibus form have been proposed by Tukey [6] and Cox [4]; the confidence curve form, illustrated below, has proved convenient in the development of some new theoretical and computational methods of estimation reported elsewhere [1].

The use of tests, especially without consideration of power, is also overly schematic for many purposes of informative inference, and there is increasing awareness that many problems customarily formulated in terms of testing can be treated more appropriately as problems of estimation. The recent paper by Natrella [5] describes this trend and some reasons for it, and illustrates how the relation between confidence intervals and tests facilitates a shift of formulations and techniques. In this connection, confidence curve estimates can be interpreted as representing critical levels and power of tests when particular hypotheses are of interest, or as incorporating testing techniques within estimation techniques when the latter are more appropriate.

Confidence curve estimates can generally be obtained by simple adaptations of standard estimation and testing techniques.

[1] Research supported by the Office of Naval Research.

Questions of justification of confidence curve estimation from the standpoint of the foundations of statistical inference lie outside the scope of the present note, but are discussed elsewhere [2], [3]. Problems of efficiency of confidence curve estimates are also discussed elsewhere [1].

2. CONFIDENCE CURVES: DEFINITION AND EXAMPLE

In problems of estimation of a real-valued parameter, a confidence curve estimate is formally defined simply as a set of upper and lower confidence limits, at each confidence coefficient from .5 to 1, inclusive; or alternatively as a set of "equal tail-area" confidence intervals, at each confidence coefficient from 0 to 1, inclusive. It is natural to require that for each possible sample, the confidence intervals are nested, those with larger coefficients including those with smaller coefficients; this restriction is easily met in practice.

In any specific application, the form and degree of completeness in which such estimates are reported can, of course, vary greatly, according to the judgment and convenience of those using the estimates. For many purposes, a rough indication, based on several confidence intervals or limits and perhaps a point estimate, may suffice; and even when a single confidence interval or limit is sufficient for some purpose of informative inference, it may be helpful to regard it as a very incomplete indication of a complete confidence curve, so as to avoid overly schematic interpretations.

For some applications a graphical representation will be convenient, and the term "confidence curve" may be used also to refer to the following specific graphical form of an estimate: For each c, $0 \leq c \leq .5$, let $\theta_L(t, c)$ and $\theta_U(t, c)$, respectively, denote lower and upper confidence limits for an unknown parameter θ, at the $1-c$ level, based on the observed value of some suitable statistic t. Such a pair of estimates also represents a $1-2c$ level confidence interval. In the (θ, c) plane, for each $c<.5$ plot the two points $(\theta_L(t, c), c)$ and $(\theta_U(t, c), c)$. For $c=.5$, we have $\theta_L(t, c) \equiv \theta_U(t, c)$, which is a median-unbiased point estimate of θ, represented by the point $(\theta_L(t, .5), .5)$. (A point estimator of θ is called median-unbiased if its probabilities of over- and under-estimation are equal, for all θ.) We denote this graph, or the function of θ which it represents, by $c(\theta, t)$. In most problems of interest, such a graph is continuous and resembles that in Figure 1.

In problems involving many familiar simple distributions, the best choice of the statistic t is the usual sufficient statistic which leads to uniformly best one-sided tests and to the corresponding standard best confidence limit estimators. (In more complicated problems, efficiency considerations lead to methods of confidence curve estimation not necessarily based on use of a single statistic t; the theory of such methods, and numerical examples, have been reported elsewhere [1].) Once a basic statistic t has been chosen, then for each possible value of t and of c, the corresponding value of $\theta_U(t, c)$ is determined by the usual condition for defining a confidence limit estimator, namely: $\theta_U(t, c)$ is that value of θ for which the $100c\%$ point of the statistic's distribution equals t. Similarly, $\theta_L(t, c)$ is the value of θ such that the $100(1-c)\%$ point equals t.

When the statistic t has discrete distributions, one may choose one of the usual alternatives: (a) when the discontinuities are small, and a convenient close continuous approximation formula is available, as in the binomial example below, the discontinuities may be ignored and the approximations may be used for typical purposes; (b) confidence limits having bounded, but not constant, confidence coefficients may be used as the elements of a confidence curve; or (c) in principle, randomization could be used to give effectively continuous distributions.

Example. Consider the problem of estimation of a binomial mean (proportion) p, based on $n=75$ observations and an observed proportion $\hat{p}=45/75=.6$. The statistic $t=\hat{p}=x/n$ has a distribution which is approximately normal, provided that np and $n(1-p)$ are not very small, namely

$$\text{Prob}(X/n \leqq v \mid p) \doteq \Phi(n^{1/2}(v-p)/(p(1-p))^{1/2}), \tag{1}$$

$$\text{where} \quad \Phi(u) = (2\pi)^{-1/2}\int_{-\infty}^{u} \exp\left(-\tfrac{1}{2}u^2\right)du.$$

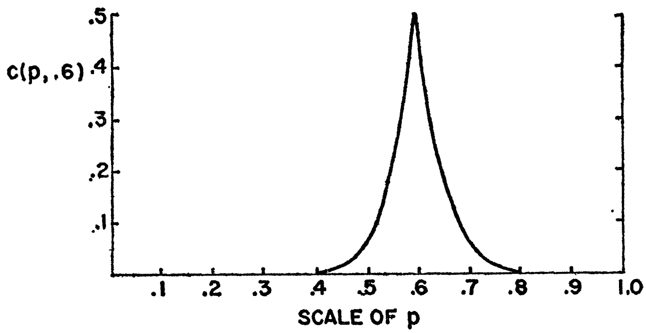

FIG. 1. Confidence curve estimate of a binomial proportion p based on $n=75$ observations and an observed proportion $\hat{p}=x/n=45/75=.6$.

For such binomial problems generally, it is easily verified that the approximate formula for a confidence curve estimate is

$$c(p, \hat{p}) = \Phi(-n^{1/2}\,|\,p-\hat{p}\,|\,/(p(1-p))^{1/2}), \tag{2}$$

the approximation being close except where np, $n(1-p)$, $n\hat{p}$, or $n(1-\hat{p})$ are very small. In the present example, we have

$$c(p, .6) = \Phi(-(8.66)\,|\,p-.6\,|\,/(p(1-p))^{1/2}). \tag{3}$$

We observe that $c(.6, .6)=.5$, and we may evaluate $c(p, .6)$ for as many additional values of p as we wish to complete a sketch of the estimate $c(p, .6)$. Such computations can, when desired, be based on the rougher approximation formula obtained by replacing $p(1-p)$ by $\hat{p}(1-\hat{p})$ in the above formula; or when convenient they may be based directly on available tables of the binomial distribution, or on available tables or graphs giving binomial confidence limits at a number of levels. A graph of the confidence curve estimate $c(p, .6)$ is given in Figure 1.

A confidence curve as a whole is an estimate; its over-all meaning and interpretation can be illustrated in part by interpretations of its various parts, which coincide with customary types of estimates and tests and their customary interpretations when used for informative inference. For example, the binomial confidence curve above admits the following partial interpretations, not all of which would be of particular interest in any single context of application. Depending upon the completeness and accuracy of plotting, such interpretations may be read more or less accurately by inspection of a graph such as Figure 1:

1. A point estimate of p is .6. (Here and in many common examples the best median-unbiased estimates obtained in this way coincide exactly or very nearly with the usual best unbiased (mean-unbiased) estimates, and with maximum likelihood estimates, except for very small sample sizes in some problems.)

2. A lower 95.5% confidence limit for p is .5.

3. A 98.4% (equal tail-area) confidence interval for p is (.45, .74).

4. For the hypothesis $p \leq .5$ (or $p \equiv .5$), against the alternative hypothesis $p > .5$, the critical level is .045 ("just significant at the 5% level"). The power of the test which rejects when the critical level is at least as small as that observed is: (a) against the alternative $p = .74 : .992 = 1 - c(.74, .6)$; (b) against the alternative $p = .57$: only $.295 = c(.57, .6)$. (When the critical level of a one-sided test is less than .5, as in cases of principal interest, then the power of the corresponding test is given as just illustrated, by $c(\theta, t)$ for alternatives θ on the same side of the maximum of the curve as the null hypothesis value(s) of θ; and by $1 - c(\theta, t)$ for alternatives on the opposite side of the maximum.)

5. For the hypothesis $p \geq .5$, the critical level is .955, which is far from suggesting rejection.

6. For the hypothesis $p = .5$, against the two-sided alternative, the critical level (of the equal tail-area test) is .09 (twice that found in the one-sided test in 4. above).

For simultaneous estimation of several parameters, analogous methods can be based upon use of nested families of confidence regions. Their graphical representation is difficult, except in the case of two parameters where the boundaries of a number of confidence regions, at selected levels, can be sketched and labeled by their confidence coefficients.

REFERENCES

[1] Birnbaum, Allan, "A unified theory of estimation, I.," *Annals of Mathematical Statistics*, 32 (1961), 112–35.

[2] Birnbaum, Allan, "On the foundations of statistical inference: Binary experiments," *Annals of Mathematical Statistics*, 32 (June, 1961).

[3] Birnbaum, Allan, "On the foundations of statistical inference. II," Technical Report IMM-NYU 275, Institute of Mathematical Sciences, New York University, 1960. To be published.

[4] Cox, D. R., "Some problems connected with statistical inference," *Annals of Mathematical Statistics*, 29 (1958), 357–72.

[5] Natrella, Mary G., "The relation between confidence intervals and tests of significance," *The American Statistician*, 14 (1960), No. 1, pp. 20–2 and 38.

[6] Tukey, John W., "Standard confidence points," Memorandum Report 26, Statistical Research Group, Princeton University, July 26, 1949.