Chapter 15

APPROXIMATING THE DISTRIBUTIONS OF ECONOMETRIC ESTIMATORS AND TEST STATISTICS

THOMAS J. ROTHENBERG*

University of California, Berkeley

Contents

*I am indebted to Christopher Cavanagh for his help in preparing this survey and to Donald Andrews and James Reeds for comments on an earlier draft. Research support from the National Science Foundation is gratefully acknowledged.

Handbook of Econometrics, Volume II, Edited by Z. Griliches and M.D. Intriligator

1. Introduction

Exact finite-sample probability distributions of estimators and test statistics are available in convenient form only for simple functions of the data and when the likelihood function is completely specified. Often in econometrics these conditions are not satisfied and inference is based on approximations to the sampling distributions. Typically "large sample" methods of approximation based on the central limit theorem are employed. For example, if $\hat{\theta}_n$ is an estimator of a parameter θ based on a sample of size n, it is sometimes possible to find a function $\sigma(\theta)$ such that the distribution of the variable $\sqrt{n}(\hat{\theta}_n - \theta)/\sigma(\theta)$ converges to a standard normal as n tends to infinity. In that case, it is common practice to approximate the distribution of $\hat{\theta}$ by a normal distribution with mean θ and variance $\sigma^2(\theta)/n$. Similar approximations are used for test statistics, although the limiting distribution is often chi-square rather than normal in this context.

These large-sample or asymptotic approximations may be quite accurate even for very small samples. The arithmetic average of independent draws from a rectangular distribution has a bell-shaped distribution for n as low as three. However, it is also easy to construct examples where the asymptotic approximation is poor even when the sample contains hundreds of observations. It is desirable, therefore, to know the conditions under which the asymptotic approximations are reasonable and to have available alternative methods when the asymptotic approximations break down. In what follows we survey some of the basic methods that have been used to approximate distributions in econometrics and describe some typical applications of these methods. Particular emphasis will be placed on "second-order" approximation methods which can be used to compare alternative asymptotically indistinguishable inference procedures.

The subject of our investigation has a long history. Techniques for approximating probability distributions have been studied by mathematical statisticians since the nineteenth century. Indeed, many of the basic methods in current use were developed more than 75 years ago. The transfer of these ideas to econometrics, however, has been very slow; only in the past 15 years has there been substantial progress in improving the approximations used in empirical economics. The reasons for this lag are not hard to fathom. The original work concentrated on one-dimensional statistics based on sums of identically distributed independent random variables. The generalization to multidimensional cases with nonlinearity, dependency, and other complications turns out to involve quite difficult mathematics and nontrivial computation. The advent of more powerful mathematical tools and enormously reduced computation cost in recent years has produced a revolution in the field of statistical approximation. Not only have old methods

been applied to more complex problems, a new burst of interest in higher-order asymptotic theory has occurred among mathematical statisticians. With so much recent development both within and without econometrics, this survey must necessarily be incomplete and tentative. It represents a somewhat personal view of the current state of a rapidly changing area of research.

Before turning to the various techniques and applications, it is perhaps useful to raise some general issues concerning the use of approximate distributions in econometrics. First of all, one must decide what one is trying to approximate. In many applications the parameter vector of interest has high dimension. Do we wish to approximate the joint probability distribution of the vector of estimates, or do we wish to approximate each marginal distribution? Is it the cumulative distribution function that needs to be approximated, or is it the density function? Some approaches which lead to good approximations of univariate densities are not convenient for obtaining good approximations of multivariate cumulative distribution functions. In practice the type of approximation method to be employed is strongly influenced by the type of function being approximated. The emphasis in the present survey will be on approximations to univariate distribution functions. It appears that most applications require knowledge of the probability that a scalar random variable lies in some interval. For example, the degree of concentration of an estimator and the power of a test can be measured by such probability statements. Although some discussion of density approximations will be presented, we shall rarely depart from distributions on the real line.

A second issue concerns the approximation of moments. If determining the full probability distribution of a statistic is hard perhaps one can get by with summary values. For many purposes, knowledge of the first few moments of an estimator or test statistic is sufficient. Thus, methods for approximating moments may be just as valuable as methods for approximating distributions. As we shall see, these methods are not unrelated: approximate moments play a key role in developing approximate distribution functions. Hence our survey will cover both topics.

Finally, and perhaps most crucially, there is the issue: What use will be made of the approximation? Generally one can distinguish two distinct reasons for wanting to know the probability distribution of an estimator or test statistic. One reason is that it is needed to make some numerical calculation from the data. For example, one might use the probability distribution to form a confidence interval for a parameter estimate; or one might form a rejection region for a test statistic. An alternative reason for knowing the probability law is that it is needed to evaluate or compare statistical procedures. One might use an estimator's probability distribution to judge whether it was reasonably accurate; or one might use sampling distributions to decide which of two tests is most powerful.

These two different uses of the probability distribution suggest different criteria for judging an approximation. For the former use, we need a computer algorithm

that will calculate, quickly and accurately, a number from the actual data. As long as the algorithm is easy to program and does not require too much data as input, it does not matter how complicated or uninterpretable it is. For the latter use, we need more than a number. The probability distribution for an estimator or test statistic generally depends on the unknown parameters and on the values of the exogenous variables. To evaluate statistical procedures we need to know how the key aspects of the distribution (center, dispersion, skewness, etc.) vary with the parameters and the exogenous data. An algorithm which computes the distribution function for any given parameter vector and data set may not be as useful as a simple formula that indicates how the shape of the distribution varies with the parameters. Interpretability, as well as accuracy, is important when comparison of probability distributions is involved.

Since my own interests are concerned with comparing alternative procedures, the present survey emphasizes approximations that yield simple analytic formulae. After reviewing a number of different approaches to approximating distributions in Section 2, the remainder of the chapter concentrates on higher-order asymptotic theory based on the Edgeworth expansion. Although the asymptotic approach rarely leads to the most accurate numerical approximations, it does lead to a powerful theory of optimal estimates and tests. In this context, it is worth recalling the words used by Edgeworth (1917) when discussing the relative merits of alternative approaches to representing empirical data: "I leave it to the impartial statistician to strike the balance between these counterpoised considerations.... I submit, too, that the decision turns partly on the purpose to which representation of statistics is directed. But one of the most difficult questions connected with our investigation is: What is its use?"

2. Alternative approximation methods

2.1. *Preliminaries*

If we are given the probability distribution of a vector of random variables, we can, in principle, find the distribution of any smooth function of these random variables by multivariate calculus. In fact, however, the mathematics is often too difficult and analytic results are unobtainable. Furthermore, we sometimes wish to learn about certain features of the distribution of a function without specifying completely the exact distribution of the underlying random variables. In this section we discuss a number of alternative methods that can be employed to obtain approximations to the probability distributions of econometric estimators and test statistics under various circumstances.

Although there is a huge statistical literature on the theory and practice of approximating distributions, there are relatively few introductory presentations of

this material. The statistics textbook by Bickel and Doksum (1977) gives a very brief survey; the handbook of distributions by Johnson and Kotz (1970) has a more comprehensive discussion. Traditional large-sample theory is developed in Cramer (1946); a detailed treatment is given in Serfling (1980). The extension to asymptotic expansions is presented in Wallace's (1958) excellent (but slightly dated) survey article; some recent developments are discussed in Bickel (1974). For a comprehensive treatment of the subject, however, a major incursion into the textbooks of advanced probability theory and numerical analysis is necessary. For those with the time and patience, chapters 15 and 16 of Feller (1971) and chapters 1, 3, and 4 of Olver (1974) are well worth the effort. In what follows we refer mostly to recent developments in the econometric literature; the bibliographies in the above-mentioned works can give entrée into the statistical literature. The recent survey paper by Phillips (1980) also gives many key references.

The present discussion is intended to be introductory and relatively nontechnical. Unfortunately, given the nature of the subject, considerable notation and formulae are still required. A few notational conventions are described here. Distribution functions will typically be denoted by the capital letters F and G; the corresponding density functions are f and g. The standard univariate normal distribution function is represented by Φ and its density by ϕ. If a p-dimensional random vector X is normally distributed with mean vector μ and covariance matrix Σ, we shall say that X is $N_p(\mu, \Sigma)$; when $p = 1$, the subscript will be dropped. The probability of an event will be indicated by $\Pr[\cdot]$. Thus, if X is $N_p(\mu, \Sigma)$ and c is a p-dimensional column vector, $\Pr[c'X \leq x] = \Phi[(x - c'\mu)/\sqrt{c'\Sigma c}]$, for all real x.

If X is a scalar random variable with distribution function F, its characteristic function is defined as $\psi(t) = \mathrm{E}\exp\{\mathrm{i}tX\}$, where t is real, E represents expectation with respect to the distribution of X, and $\mathrm{i} = \sqrt{-1}$. The function $K(t) = \log \psi(t)$ is called the cumulant function. If X possesses moments up to order r, then $\psi(t)$ is differentiable up to order r; furthermore, the rth moment of X is given by the rth derivative of $\mathrm{i}^{-r}\psi(t)$ evaluated at zero:

$$\mathrm{E}(X^r) = \mathrm{i}^{-r}\psi^{(r)}(0).$$

The rth derivative of $\mathrm{i}^{-r}K(t)$, evaluated at zero, is called the rth cumulant of X and is denoted by:

$$k_r = \mathrm{i}^{-r}K^{(r)}(0).$$

Since the derivatives of $K(t)$ are related to the derivatives of $\psi(t)$, the cumulants are related to the moments. In fact, k_1 is the mean and k_2 is the variance. For a standardized random variable with zero mean and unit variance, k_3 is the third moment and k_4 is the fourth moment less three. For a normal random variable,

all cumulants of order greater than two are zero. Hence, these cumulants can be viewed as measures of departure from normality. For further details, one may consult Kendall and Stuart (1969, ch. 3).

Our discussion will concentrate on approximating the cumulative distribution functions for continuous random variables. If the approximating distribution function is differentiable, there will generally be no problem in obtaining an approximate density function. Some approximation methods, however, apply most easily to the density function directly. In that case, numerical integration may be needed to obtain the distribution function if analytic integration is difficult.

2.2. *Curve-fitting*

The simplest way to approximate a distribution is to find a family of curves possessing the right shape and select that member which seems to fit best. If the low-order moments of the true distribution are known, they can be used in the fitting process. If not, Monte Carlo simulations or other information about the true distribution can be employed instead.

Durbin and Watson (1971) describe a number of different approximations to the null distribution of their d statistic for testing serial correlation in regression disturbances. One of the most accurate is the beta approximation proposed by Henshaw (1966). Since d must lie between zero and four and seems to have a unimodal density, it is not unreasonable to think that a linear transformed beta distribution might be a good approximation to the true distribution. Suppose X is a random variable having the beta distribution function:

$$\Pr[X \leq x] = \int_0^x \frac{1}{B(p,q)} t^{p-1}(1-t)^{q-1}\,\mathrm{d}t \equiv G(x; p, q).$$

Then, for constants a and b, the random variable $a + bX$ has moments depending on p, q, a, and b. These moments are easy to express in analytic form. Furthermore, the moments of the Durbin–Watson statistic d are also simple functions of the matrix of regression variables. Equating the first four moments of d to the corresponding moments of $a + bX$, one obtains four equations in the four parameters. For any given matrix of observations on the regressors, these equations give unique solutions, say p^*, q^*, a^*, b^*. Then $\Pr[d \leq x]$ can be approximated by $G[(x - a^*)/b^*; p^*, q^*]$. This approximation appears to give third decimal accuracy for a wide range of cases. Theil and Nagar (1961) had earlier proposed a similar approximation, but used approximate rather than actual moments of d. Since these approximate moments do not vary with the matrix of regressors, the Theil–Nagar approximation is independent of the data and can be

tabulated once and for all. Unfortunately, the moment approximation is not always accurate and the resulting approximation to the probability distribution is less satisfactory than Henshaw's.

A more sophisticated version of the curve-fitting method is suggested by Phillips (1981). Suppose a statistic X is known to have a density function $f(x)$ that behaves in the tails like the function $s(x)$. For example, if X possess moments only up to order k and takes values everywhere on the real line, $f(x)$ might behave in the tails like a Student density with $k+1$ degrees of freedom. For some small integer r, one might approximate the density function $f(x)$ by a rational function modification of $s(x)$:

$$s(x)\frac{a_0 + a_1 x + \cdots + a_r x^r}{b_0 + b_1 x + \cdots + b_r x^r}, \tag{2.1}$$

where the a_i and b_i are chosen to make the approximation as accurate as possible. Since the function (2.1) does not typically have simple moment formulae (or even possess finite moments), the method of moments is not a useful way to obtain values for the a_i and b_i. But, Monte Carlo experimental data or local power series expansions of the density may be available to help select the parameters. Since (2.1) has $2r+1$ free parameters, it appears that, with a judicious choice for s, this functional form should provide a very accurate approximation to the density function of any econometric statistic. Furthermore, if s is replaced by its integral, a function of the same form as (2.1) could be used to approximate a distribution function.

If considerable information about the true density is available, curve-fitting methods are likely to provide simple and very accurate approximations. Phillips (1981) produces some striking examples. Indeed it is unlikely that any other method will give better numerical results. However, curve-fitting methods are considerably less attractive when the purpose is not quantitative but qualitative. Comparisons of alternative procedures and sensitivity analysis are hindered by the fact that curve-fitting methods do not typically yield a common parametric family. If two statistics are both (approximately) normal, they can be compared by their means and variances. If one statistic is approximately beta and the other approximately normal, comparisons are difficult: the parameters that naturally describe one distribution are not very informative about the other. The very flexibility that makes curve-fitting so accurate also makes it unsuitable for comparisons.

2.3. *Transformations*

Suppose X is a random variable and h is a monotonically increasing function such that $h(X)$ has a distribution function well approximated by G. Since $\Pr[X \leq x]$ is

the same as $\Pr[h(X) \leq h(x)]$, the distribution function for X should be well approximated by $G[h(x)]$. For example, if X has a chi-square distribution with k degrees of freedom, $\sqrt{X} - \sqrt{k}$ has approximately a $N(0, \frac{1}{2})$ distribution when k is large. Hence, one might approximate $\Pr[X \leq x]$ by $\Phi[\sqrt{2x} - \sqrt{2k}]$. A better approach, due to Wilson and Hilferty (1931), is to treat $(X/k)^{1/3}$ as $N(1, 2/9k)$ and to approximate $\Pr[X \leq x]$ by $\Phi[((x/k)^{1/3} - 1)\sqrt{9k/2}]$.

Fisher's z transformation is another well-known example of this technique. The sample correlation coefficient $\hat{\rho}$ based on random sampling from a bivariate normal distribution is highly skewed if the population coefficient ρ is large in absolute value. However, $z = h(\hat{\rho}) = \log(1 - \hat{\rho})/(1 + \hat{\rho})$ has rather little skewness and is well approximated by a normal random variable with mean $\log(1-\rho)/(1 + \rho)$ and variance n^{-1}. Thus, $\Pr[\hat{\rho} \leq x]$ can be approximated by $\Phi[\sqrt{n}\, h(x) - \sqrt{n}\, h(\rho)]$ for moderate sample size n.

Using transformations to approximate distributions is an art. Sometimes, as in the correlation coefficient case, the geometry of the problem suggests the appropriate transformation h. Since $\hat{\rho}$ can be interpreted as the cosine of the angle between two normal random vectors, an inverse trigonometric transformation is suggested. In other cases, arguments based on approximate moments are useful. Suppose $h(X)$ can be expanded in a power series around the point $\mu = \mathrm{E}(X)$:

$$h(X) = h(\mu) + h'(\mu)(X - \mu) + \tfrac{1}{2}h''(\mu)(X - \mu)^2 + \cdots, \tag{2.2}$$

where $X - \mu$ is in some sense small.[1] Then we might act as though:

$$\mathrm{E}(h) \approx h(\mu) + \tfrac{1}{2}h''(\mu)\mathrm{E}(X - \mu)^2,$$

$$\mathrm{Var}(h) \approx \left[h'(\mu)\right]^2 \mathrm{Var}(X),$$

$$\mathrm{E}(h - \mathrm{E}h)^3 \approx \left[h'(\mu)\right]^3 \mathrm{E}(X - \mu)^3 + \tfrac{3}{2}\left[h'(\mu)\right]^2 h''(\mu)\left[\mathrm{E}(X - \mu)^4 - \mathrm{Var}^2(X)\right],$$

and choose h so that these approximate moments match the moments of the approximating distribution. If the approximating distribution is chosen to be normal, we might require that $\mathrm{Var}(h)$ be a constant independent of μ; or we might want the third moment to be zero. If the moments of X are (approximately) known and the above approximations used, either criterion gives rise to a differential equation in $h(\mu)$. The cube-root transformation for the chi-square random variable can be motivated on the grounds it makes the approximate third moment of h equal to zero. The Fisher transformation for $\hat{\rho}$ stabilizes the approximate variance of h so that it is independent of ρ.

[1] For example, if X is a statistic from a sample of size n, its variance might be proportional to n^{-1}. Expansions like (2.2) are discussed in detail in Section 3 below.

Transformations are discussed in detail by Johnson (1949) and illustrated by numerous examples in Johnson and Kotz (1970). Jenkins (1954) and Quenouille (1948) apply inverse trigonometric transformations to the case of time-series autocorrelation coefficients. The use of transformations in econometrics, however, seems minimal, probably because the method is well developed only for univariate distributions. Nevertheless, as an approach to approximating highly skewed distributions, transformations undoubtedly merit further study.

2.4. *Asymptotic expansions*

Often it is possible to embed the distribution problem at hand in a sequence of similar problems. If the sequence has a limit which is easy to solve, one might approximate the solution of the original problem by the solution of the limit problem. The sequence of problems is indexed by a parameter which, in many econometric applications, is the sample size n. Suppose, for example, one wishes to approximate the probability distribution of an estimator of a parameter θ based on a sample. We define an infinite sequence $\hat{\theta}_n$ of such estimators, one for each sample size $n = 1, 2, \ldots$, and consider the problem of deriving the distribution of each $\hat{\theta}_n$. Of course, we must also describe the joint probability distribution of the underlying data for each n. Given such a sequence of problems, the asymptotic approach involves three steps: (a) A simple monotonic transformation $T_n = h(\hat{\theta}_n; \theta, n)$ is found so that the distribution of the transformed estimator T_n is not very sensitive to the value n. Since most interesting estimators are centered at the true parameter and have dispersion declining at the rate $n^{-1/2}$, the linear transformation $T_n = \sqrt{n}(\hat{\theta}_n - \theta)$ is often used. (b) An approximation $G_n(x)$ to the distribution function $F_n(x) \equiv \Pr[T_n \leq x]$ is found so that, as n tends to infinity, the error $|G_n(x) - F_n(x)|$ goes to zero. (c) The distribution function for $\hat{\theta}_n$ is approximated using G_n; that is, $\Pr[\hat{\theta}_n \leq t] \equiv \Pr[T_n \leq h(t; \theta, n)]$ is approximated by $G_n[h(t; \theta, n)]$.

For many econometric estimators $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal. Hence, using the linear transformation for h, one may choose a normal distribution function for G_n. However, it is possible to develop other approximations. Let $G_n(x)$ be an approximation to the continuous distribution function $F_n(x)$. If, for all x,

$$\lim_{n \to \infty} n^r |F_n(x) - G_n(x)| = 0,$$

we write

$$F_n(x) = G_n(x) + o(n^{-r})$$

and say that G_n is a $o(n^{-r})$ approximation to F_n. (A similar language can be developed for approximating density functions.) The asymptotic distribution is a $o(n^0)$ approximation.

The number r measures the speed at which the approximation error goes to zero as n approaches infinity. Of course, for given sample size n, the value of r does not tell us anything about the goodness of the approximation. If, however, we have chosen the transformation h cleverly so that F_n and G_n vary smoothly with n, the value of r might well be a useful indicator of the approximation error for moderate values of n.

There are two well-known methods for obtaining higher-order approximate distribution functions based on Fourier inversion of the approximate characteristic function. Let $\psi(t) = \mathrm{E}\exp\{\mathrm{i}tT_n\}$ be the characteristic function for T_n and let $K(t) \equiv \log \psi(t)$ be its cumulant function. If ψ is integrable, the density function f_n for T_n can be written as:[2]

$$f_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}xt}\psi(t)\,\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}xt+K(t)}\,\mathrm{d}t. \tag{2.3}$$

Often $K(t)$ can be expanded in a series where the successive terms are increasing powers of $n^{-1/2}$. The integrand can then be approximated by keeping only the first few terms of the series expansion. Integrating term by term, one obtains a series approximation to f_n; further integration yields a series approximation to the distribution function. The Edgeworth approximation (which is obtained by expanding $K(t)$ around $t = 0$) is the simplest and most common method; it does not require complete knowledge of $K(t)$ and can be calculated from the low-order cumulants of T_n. A detailed discussion of the Edgeworth approximation appears in Section 3. The saddlepoint approximation (which is obtained by expanding $K(t)$ around the "saddlepoint" value t^* that maximizes the integrand) is more complex and requires intimate knowledge of the cumulant function. When available, it typically gives more accurate approximations especially in the tails of the distribution. Daniels (1956) and Phillips (1978) have applied the method to some autocorrelation statistics in time series. Unfortunately, knowledge of the cumulant function is rare in econometric applications; the saddlepoint approximation has therefore received little attention to date and will not be emphasized in the present survey.

Wallace (1958) presents an excellent introduction to asymptotic approximations based on expansions of the characteristic function. An exposition with emphasis on multivariate expansions is given by Barndorff-Nielsen and Cox (1979); the comments on this paper by Durbin (1979) are particularly interesting and suggest new applications of the saddlepoint method. In econometrics, T. W.

[2] Cf. Feller (1981, p. 482).

Anderson, P. C. B. Phillips, and J. D. Sargan have been pioneers in applying asymptotic expansions. Some of their work on estimators and test statistics in simultaneous equations models is surveyed in Section 6 below.

The above discussion of asymptotic expansions has focused on estimators, but there is no difficulty in applying the methods to any sample statistic whose cumulant function can be approximated by a power series in $n^{-1/2}$. Furthermore, the parameter which indexes the sequence of problems need not be the sample size. In the context of the simultaneous equations model, Kadane (1971) suggested that it might be more natural to consider a sequence indexed by the error variance. In his "small σ" analysis, the reduced-form error-covariance matrix is written as $\sigma\Omega$; in the sequence, the sample size and the matrix Ω are fixed, but σ approaches zero. Edgeworth and saddlepoint expansions are available as long as one can expand the cumulant function in a power series where successive terms are increasing powers of σ. Anderson (1977) explores this point of view in the context of single-equation estimation in structural models.

2.5. *Ad hoc methods*

Certain statistics permit approximations which take advantage of their special structure. Consider, for example, the ratio of two random variables, say $T = X_1/X_2$. If X_2 takes on only positive values, $\Pr[T \leq x] = \Pr[X_1 - xX_2 \leq 0]$. If both X_1 and X_2 are sums of independent random variables possessing finite variance, then the distribution of $X_1 - xX_2$ might be approximated by a normal. Defining $\mu_i = \mathrm{E}X_i$ and $\sigma_{ij} = \mathrm{E}(X_i - \mu_i)(X_j - \mu_j)$, we might approximate $\Pr[T \leq x]$ by:

$$\Phi\left[(x\mu_2 - \mu_1)\Big/\sqrt{\sigma_{11} - 2\sigma_{12}x + \sigma_{22}x^2}\right].$$

Even if X_2 is not always positive, as long as $\Pr[X_2 \leq 0]$ is negligible, the above approximation might be reasonable.

An important example of this situation occurs when X_1 and X_2 are quadratic forms in normal variables. Suppose $X_1 = z'Az$ and $X_2 = z'Bz$, where z is $N_p(0, \Sigma)$. Then, by a rotation of the coordinate system, $X \equiv X_1 - xX_2$ can be written as:

$$X = z'(A - xB)z = \sum_{i=1}^{p} \lambda_i y_i^2,$$

where the λ_i are the characteristic roots of $\Sigma(A - xB)$ and the y_i are independent $N(0,1)$ random variables. If p is moderately large (say, 20 or more) and the λ_i are not too dispersed, a central limit theorem might be invoked and the distribution of X approximated by a normal with mean $\mathrm{tr}\,\Sigma(A - xB)$ and

variance $2\,\mathrm{tr}[\Sigma(A - xB)]^2$. If necessary, Edgeworth or saddlepoint expansions (in powers of $p^{-1/2}$) could be employed to obtain greater accuracy.

In this quadratic case, approximations can be dispensed with entirely. The exact distribution of a weighted sum of independent chi-square random variables can be obtained by one-dimensional numerical integration using the algorithms of Imhof (1961) or Pan Jie-jian (1968). Koerts and Abrahamse (1969) and Phillips (1977a), among others, have used these methods to calculate exact distributions of some time-series statistics. For large p, numerical integration is unnecessary since the Edgeworth approximation to X is likely to be adequate. [Cf. Anderson and Sawa (1979).]

The least-squares estimator of a single coefficient in a linear regression equation can always be written as a ratio. In particular, when one of the regressors is endogenous, its coefficient estimator is the ratio of two quadratic forms in the endogenous variables. Thus ratios occur often in econometrics and their simple structure can easily be exploited. The multivariate generalization $X_2^{-1}X_1$, where X_2 is a random square matrix and X_1 is a random vector, also has a simple structure, but approximation methods for this case seem not to have been explored.

In practice, ad hoc techniques which take advantage of the special structure of the problem are invaluable for developing simple approximations. General methods with universal validity have attractive theoretical features, but are not particularly accurate for any given problem. Approximating distributions is an art involving judgment and common sense, as well as technical skill. The methods discussed in this section are not distinct alternatives. Every approximation involves fitting a curve to a transformed statistic, dropping terms which are judged to be small. In the end, many approaches are merged in an attempt to find a reasonable solution to the problem at hand.

3. Edgeworth approximations

Perhaps the most important and commonly used method to obtain improved approximations to the distributions of estimators and test statistics in econometrics is the Edgeworth expansion. There are a number of reasons for this prominence. First, the method is a natural extension of traditional large-sample techniques based on the central limit theorem. The usual asymptotic approximation is just the leading term in the Edgeworth expansion. Second, since the expansion is based on the normal and chi-square distributions—which are familiar and well tabulated—it is easy to use. Finally, the method can be employed to approximate the distributions of most of the commonly used estimators and test statistics and is very convenient for comparing alternative statistical procedures. Indeed, it is the basis for a general theory of higher-order efficient estimators and tests.

Because of its prominence, the Edgeworth approximation will be described at some length in this section and in the examples which follow. However, it is worth noting at the outset that Edgeworth methods do not lead to particularly accurate approximations. To the contrary, in nearly every application, there exist alternative curve-fitting techniques yielding more satisfactory numerical results. Edgeworth is important, not for its accuracy, but for its general availability and simplicity. Although rarely optimal, it is often quite adequate and leads to a useful, comprehensive approach to second-order comparisons of alternative procedures.

Our discussion of the Edgeworth expansion parallels the traditional approach to asymptotic distribution theory as presented in Theil (1971, ch. 8) or Bishop, Fienberg, and Holland (1975, ch. 14). We first consider the problem of approximating sums of independent random variables. Then we show that the theory also applies to smooth functions of such sample sums. To avoid excessive length and heavy mathematics, our presentation will be quite informal; rigorous proofs and algebraic detail can be found in the literature cited. Although Edgeworth expansions to high order are often available, in practice one rarely goes beyond the first few terms. We shall develop the expansion only up to terms of order n^{-1} and refer to the result as the "second-order" or $o(n^{-1})$ Edgeworth approximation. The extension to higher terms is in principle straightforward, but the algebra quickly becomes extremely tedious.

3.1. *Sums of independent random variables*

Suppose $X_1, X_2, \ldots$ form an infinite sequence of independent random variables with common density function f; each X_i has mean zero, variance one, and possesses moments up to the fourth order. If ψ is the characteristic function associated with f, then the cumulant function $\log \psi$ possesses derivatives up to the fourth order and can be expanded in a neighborhood of the origin as a power series:

$$\log \psi(t) = \tfrac{1}{2}(it)^2 + \tfrac{1}{6}k_3(it)^3 + \tfrac{1}{24}k_4(it)^4 + \cdots, \tag{3.1}$$

where k_r is the rth cumulant of f.

The standardized sum $T_n = \sum X_i/\sqrt{n}$ also has mean zero and variance one; let f_n and ψ_n be its density and characteristic functions. Since

$$\begin{aligned}\log \psi_n(t) &= n \log \psi\left(t/\sqrt{n}\right) \\ &= \tfrac{1}{2}(it)^2 + \frac{1}{6\sqrt{n}}k_3(it)^3 + \frac{1}{24n}k_4(it)^4 + \cdots,\end{aligned}$$

we observe that the rth cumulant of T_n is simply $k_r n^{1-r/2}$, for $r > 2$. Thus, the high-order cumulants are small when n is large. Since the function e^x has the expansion $1 + x + \frac{1}{2}x^2 + \cdots$, when x is small, $\psi_n(t)$ can be written as a power series in $n^{-1/2}$:

$$\begin{aligned}\psi_n(t) &= \exp\{\log \psi_n(t)\} \\ &= e^{-1/2t^2}\left[1 + \frac{1}{6\sqrt{n}} k_3 (it)^3 + \frac{3k_4 (it)^4 + k_3^2 (it)^6}{72n} + \cdots\right].\end{aligned} \tag{3.2}$$

The $o(n^{-1})$ Edgeworth approximation to the density function for T_n is obtained by applying the Fourier inversion formula (2.3) and dropping high-order terms. Using the fact that, if f has characteristic function $\psi(t)$, then the rth derivative $f^{(r)}$ has characteristic function $(-it)^r \psi(t)$, the inverse Fourier transform is seen to be:

$$\begin{aligned}f(x) &\approx \varphi(x) - \frac{1}{6\sqrt{n}} k_3 \varphi^{(3)}(x) + \frac{1}{24n} k_4 \varphi^{(4)}(x) + \frac{1}{72n} k_3^2 \varphi^{(6)}(x) \\ &\approx \varphi(x)\left[1 + \frac{k_3 H_3(x)}{6\sqrt{n}} + \frac{3k_4 H_4(x) + k_3^2 H_6(x)}{72n}\right],\end{aligned} \tag{3.3}$$

where $\varphi^{(r)}$ is the rth derivative of the normal density function φ and H_r is the Hermite polynomial of degree r defined as:

$$H_r(x) = (-1)^r \frac{\varphi^{(r)}(x)}{\varphi(x)}.$$

(By simple calculation, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, etc.) Integration of (3.3) gives an approximation for the distribution function:

$$F_n(x) \approx \Phi(x) - \varphi(x)\left[\frac{k_3 H_2(x)}{6\sqrt{n}} + \frac{3k_4 H_3(x) + k_3^2 H_5(x)}{72n}\right]. \tag{3.4}$$

This latter formula can be rewritten as:

$$F_n(x) \approx \Phi\left[x - \frac{k_3(x^2 - 1)}{6\sqrt{n}} + \frac{3k_4(3x - x^3) + 2k_3^2(4x^3 - 7x)}{72n}\right], \tag{3.5}$$

by use of Taylor series expansion and the definition of the Hermite polynomials.

Equations (3.4) and (3.5) are two variants of the $o(n^{-1})$ Edgeworth approximation to the distribution function of T_n; Phillips (1978) refers to them as the Edgeworth-A and Edgeworth-B approximations, respectively. The latter is closely related to the Cornish–Fisher (1937) normalizing expansion for a sample statistic. If (3.5) is written as:

$$\Pr[T_n \leq x] \approx \Phi\left[x + \frac{g_1(x)}{\sqrt{n}} + \frac{g_2(x)}{n}\right],$$

then, when n is large enough to ensure that the function in brackets is monotonic for the x values of interest, it can be rewritten as:

$$\Pr\left[T_n + \frac{g_1(T_n)}{\sqrt{n}} + \frac{g_2(T_n)}{n} \leq x\right] \approx \Phi[x].$$

Thus, the function inside brackets in (3.5) can be viewed as a transformation h, making $h(T_n)$ approximately normal.

The argument sketched above is, of course, purely formal. There is no guarantee that the remainder terms dropped in the manipulations are really small. However, with a little care, one can indeed prove that the Edgeworth approximations are valid asymptotic expansions. Suppose the power series expansion of ψ_n is carried out to higher order and $G_n^r(x)$ is the analogous expression to (3.4) when terms up to order $n^{-r/2}$ are kept. Then, if the X_i possess moments to order $r+2$ and $|\psi(t)|$ is bounded away from one for large t:

$$\Pr[T_n \leq x] = G_n^r(x) + o(n^{-r/2}).$$

A proof can be found in Feller (1971, pp. 538–542). The assumption on the characteristic function rules out discrete random variables like the binomial. Since the distribution of a standardized sum of discrete random variables generally has jumps of height $n^{-1/2}$, it is not surprising that it cannot be closely approximated by a continuous function like (3.4). Edgeworth-type approximations for discrete random variables are developed in Bhattacharya and Rao (1976), but will not be described further in the present survey.

The theory for sums of independent, identically distributed random variables is easily generalized to the case of weighted sums. Furthermore, a certain degree of dependence among the summands can be allowed. Under the same type of regularity conditions needed to guarantee the validity of a central limit theorem, it is possible to show that the distribution functions for standardized sample moments of continuous random variables possess valid Edgeworth expansions as long as higher-order population moments exist.

3.2. *A general expansion*

Since few econometric estimators or test statistics are simple sums of random variables, these classical asymptotic expansions are not directly applicable. Nevertheless, just as the delta method[3] can be applied to obtain first-order asymptotic distributions for smooth functions of random variables satisfying a central limit theorem, a generalized delta method can be employed for higher-order expansions. With simple modifications, the classical formulae (3.3)–(3.5) are valid for most econometric statistics possessing limiting normal distributions.

Nagar (1959) noted that k-class estimators in simultaneous equations models can be expanded in formal series where the successive terms are increasing powers of $n^{-1/2}$. The expansions are essentially multivariate versions of (2.2). The rth term takes the form $C_r n^{-r/2}$, where C_r is a polynomial in random variables with bounded moments. His approach is to keep the first few terms in the expansion and to calculate the moments of the truncated series. These moments can be interpreted as the moments of a statistic which serves to approximate the estimator. [In some circumstances, these moments can be interpreted as approximations to the actual moments of the estimator; see, for example, Sargan (1974).]

Nagar's approach is quite generally available and can be used to develop higher-order Edgeworth approximations. Most econometric estimators and test statistics, after suitable standardization so that the center and dispersion of the distributions are stabilized, can be expanded in a power series in $n^{-1/2}$ with coefficients that are well behaved random variables. Suppose, for example, T_n is a standardized statistic possessing the stochastic expansion:

$$T_n = X_n + \frac{A_n}{\sqrt{n}} + \frac{B_n}{n} + \frac{R_n}{n\sqrt{n}}, \tag{3.6}$$

where X_n, A_n, and B_n are sequences of random variables with limiting distributions as n tends to infinity. If R_n is stochastically bounded,[4] the limiting distribution of T_n is the same as the limiting distribution of X_n. It is natural to use the information in A_n and B_n to obtain a better approximation to the distribution of T_n. Suppose the limiting distribution of X_n is $N(0,1)$. Let $T' = X_n + A_n n^{-1/2} + B_n n^{-1}$ be the first three terms of the stochastic expansion of T_n. For a large class of cases, T' has finite moments up to high order and its rth cumulant is of order

[3]Suppose a standardized sample mean $X_n = \sqrt{n}(\bar{x} - \mu)/\sigma$ is asymptotically $N(0,1)$ and g is a differentiable function with derivative $b \equiv g'(\mu)$. The delta method exploits the fact that, when n is large, $T_n = \sqrt{n}[g(\bar{x}) - g(\mu)]$ behaves like $b\sigma X_n$; hence T_n is asymptotically $N(0, b^2\sigma^2)$. Cf. Theil (1971, pp. 373–374).

[4]A sequence of random variables Z_n is stochastically bounded if, for every $\varepsilon > 0$, there exists a constant c such that $\Pr[|Z_n| > c] < \varepsilon$, for sufficiently large n. That is, the distribution function does not drift off to infinity. Cf. Feller (1971, p. 247).

$n^{(2-r)/2}$ when r is greater than 2. Furthermore, its mean and variance can be written as:

$$\mathrm{E}(T') = \frac{a}{\sqrt{n}} + o(n^{-1}),$$

$$\mathrm{Var}(T') = 1 + \frac{b}{n} + o(n^{-1}),$$

where a and b depend on the moments of X_n, A_n, and B_n. The restandardized variable,

$$T^* = \frac{T' - \dfrac{a}{\sqrt{n}}}{\sqrt{1 + b/n}},$$

has, to order n^{-1}, zero mean and unit variance. Its third and fourth moments are:

$$\mathrm{E}(T^*)^3 = \frac{c}{\sqrt{n}} + o(n^{-1}),$$

$$\mathrm{E}(T^*)^4 = 3 + \frac{d}{n} + o(n^{-1}),$$

where $c/\sqrt{n}$ is the approximate third cumulant of T' and d/n is the approximate fourth cumulant. Since the cumulants of T' behave like the cumulants of a standardized sum of independent random variables, one is tempted to use the Edgeworth formulae to approximate its distribution. For example, one might approximate $\Pr[T^* \le x]$ by (3.5) with c replacing k_3 and d replacing k_4. Dropping the remainder term and using the fact that

$$\Pr[T' \le x] = \Pr\left[T^* \le \frac{x - a/\sqrt{n}}{\sqrt{1 + b/n}}\right]$$

$$= \Pr\left[T^* \le x - \frac{a}{\sqrt{n}} - \frac{bx}{2n} + o(n^{-1})\right],$$

we are led to the Edgeworth-B approximation:

$$\Pr[T_n \le x] \approx \Phi\left[x + \frac{\gamma_1 + \gamma_2 x^2}{6\sqrt{n}} + \frac{\gamma_3 x + \gamma_4 x^3}{72n}\right], \tag{3.7}$$

where

$$\gamma_1 = c - 6a; \qquad \gamma_3 = 9d - 14c^2 - 36b + 24ac,$$

$$\gamma_2 = -c; \qquad \gamma_4 = 8c^2 - 3d.$$

A similar calculation using (3.4) leads to an Edgeworth-A form of the approximation.

Of course, the above discussion in no way constitutes a proof that the approximation (3.7) has an error $o(n^{-1})$. We have dropped the remainder term $R_n/n\sqrt{n}$ without justification; and we have used the Edgeworth formulae despite the fact that T' is not the sum of n independent random variables. With some additional assumptions, however, such a proof can in fact be constructed.

If $|T_n - T'|$ is stochastically of order $o(n^{-1})$, it is reasonable to suppose that the distribution functions for T_n and T' differ by that order. Actually, further assumptions on the tail behavior of R_n are required. Using a simple geometric argument, Sargan and Mikhail (1971) show that, for all x and ε,

$$|\Pr(T_n \le x) - \Pr(T' \le x)| \le \Pr[|T_n - T'| > \varepsilon] + \Pr[|T' - x| < \varepsilon].$$

If T' has a bounded density, the last term is of order ε, as ε approaches zero. To show that the difference between the two distribution functions is $o(n^{-1})$ we choose ε to be of that order. Setting $\varepsilon = n^{-3/2}\log^c n$, we find that a sufficient condition for validly ignoring the remainder term is that there exists a positive constant c such that:

$$\Pr[|R_n| > \log^c n] = o(n^{-1}). \tag{3.8}$$

That is, the tail probability of R_n must be well behaved as n approaches infinity. If R_n is bounded by a polynomial in normal random variables, (3.8) is necessarily satisfied.

To show that T' can be approximated by the Edgeworth formulae, one must make strong assumptions about the sequences X_n, A_n, and B_n. If A_n and B_n are polynomials in variables which, along with X_n, possess valid Edgeworth expansions to order n^{-1}, the results of Chibisov (1980) can be used to prove a validity theorem. The special case where (3.6) comes from the Taylor series expansion of a smooth function $g(p)$, where p is a vector of sample moments, has been studied by Bhattacharya and Ghosh (1978), Phillips (1977b), and Sargan (1975b, 1976). These authors give formal proofs of the validity of the Edgeworth approximation under various assumptions on the function g and the distribution of p. Sargan (1976) gives explicit formulae for the γ_i of (3.7) in terms of the derivatives of the function g and the cumulants of p.

It may be useful to illustrate the approach by a simple example. Suppose $\bar{x}$ and s^2 are the sample mean and (bias adjusted) sample variance based on n independent draws from a $N(\mu, \sigma^2)$ distribution. We shall find the Edgeworth approximation to the distribution of the statistic:

$$T_n = \frac{\sqrt{n}(\bar{x} - \mu)}{s},$$

which, of course, is distributed exactly as Student's t. With $X_n = \sqrt{n}(\bar{x} - \mu)/\sigma$ and $Y_n = \sqrt{n}(s^2 - \sigma^2)/\sigma^2$, the statistic can be written as:

$$T_n = \frac{X_n}{\sqrt{1 + Y_n/\sqrt{n}}} = X_n - \frac{X_n Y_n}{2\sqrt{n}} + \frac{3X_n Y_n^2}{8n} + \frac{R_n}{n\sqrt{n}},$$

where the remainder term R_n is stochastically bounded. The random variable X_n is $N(0,1)$; Y_n is independent of X_n with mean zero and variance $2n/(n-1)$. It is easy to verify that T_n satisfies the assumptions of Sargan (1976) and hence can be approximated by a valid Edgeworth expansion. Dropping the remainder term, we find that T' has mean zero and variance $1 + 2n^{-1} + o(n^{-1})$. Its third cumulant is exactly zero and its fourth cumulant is approximately $6n^{-1}$. Thus, with $a = c = 0$, $b = 2$, and $d = 6$, (3.7) becomes:

$$\Pr[T_n \leq x] \approx \Phi\left[x - \frac{x + x^3}{4n}\right], \tag{3.9}$$

which is a well-known approximation to the Student-t distribution function.

There are available a number of alternative algorithms for calculating Edgeworth expansions. The use of (3.7) with Nagar-type approximate moments is often the simplest. Sometimes, however, the moment calculations are tedious and other methods are more convenient. If, for example, the exact characteristic function for T_n is known, it can directly be expanded in a power series without the need to calculate moments. The Edgeworth approximation can be found by Fourier inversion of the first few terms of the series. Anderson and Sawa (1973) employ this method in their paper on the distribution of k-class estimators in simultaneous equations models.

An alternative approach, used by Hodges and Lehmann (1967), Albers (1978), Anderson (1974), and Sargan and Mikhail (1971), exploits the properties of the normal distribution. Suppose the stochastic expansion (3.6) can be written as:

$$T_n = X_n + \frac{A(X_n, Y_n)}{\sqrt{n}} + \frac{B(X_n, Y_n)}{n} + \frac{R_n}{n\sqrt{n}},$$

where R_n satisfies (3.8), X_n is exactly $N(0,1)$, and the vector Y_n is independent of X_n with bounded moments. The functions A and B are assumed to be smooth in both arguments with A' denoting the derivative of A with respect to X_n. Then, conditioning on Y_n and supressing the subscripts, we write:

$$\begin{aligned}\Pr[T' \leq x] &= \mathrm{E}_Y \Pr\left[X + \frac{A(X,Y)}{\sqrt{n}} + \frac{B(X,Y)}{n} \leq x \mid Y\right] \\ &\approx \mathrm{E}_Y \Phi\left[x - \frac{A(x,Y)}{\sqrt{n}} - \frac{B(x,Y) - A(x,Y)A'(x,Y)}{n}\right].\end{aligned}$$

The approximation comes from dropping terms of higher order when inverting the inequality. Taking expectation of the Taylor series expansion of Φ, we obtain the approximation:

$$\Pr[T_n \leq x] \approx \Phi\left[x - \mathrm{E}_Y \frac{A(x,Y)}{\sqrt{n}} - \mathrm{E}_Y \frac{B(x,Y) - A(x,Y)A'(x,Y) + \frac{1}{2}x\mathrm{Var}_Y A(x,Y)}{n}\right]. \quad (3.10)$$

Of course, some delicate arguments are needed to show that the error of approximation is $o(n^{-1})$; some conditions on the functions A and B are clearly necessary. Typically, A and B are polynomials and the expectations involved in (3.10) are easy to evaluate. In our Student-t example, we find from elementary calculation $\mathrm{E}_Y(A) = 0$, $\mathrm{E}_Y(B) = 3x/4$, $\mathrm{E}_Y(AA') = x/2$, and $\mathrm{Var}_Y(A) = x^2/2$; hence, we obtain the approximation (3.8) once again.

3.3. *Non-normal expansions*

Edgeworth approximations are not restricted to statistics possessing limiting normal distributions. In the case of multivariate test statistics, the limiting distribution is typically chi-square and asymptotic expansions are based on that distribution. The following general algorithm is developed by Cavanagh (1983). Suppose the sample statistic T_n can be expanded as:

$$T_n = X_n + \frac{A_n}{\sqrt{n}} + \frac{B_n}{n} + \frac{R_n}{n\sqrt{n}},$$

where X_n has, to order n^{-1}, the distribution function F and density function f; the random variables A_n and B_n are stochastically bounded with conditional moments:

$$a(x) = \mathrm{E}(A_n | X_n = x),$$
$$b(x) = \mathrm{E}(B_n | X_n = x),$$
$$v(x) = \mathrm{Var}(A_n | X_n = x),$$

that are smooth functions of x. Define the derivative functions $a' = \mathrm{d}a/\mathrm{d}x$, $v' = \mathrm{d}v/\mathrm{d}x$ and

$$c(x) = \frac{\mathrm{d}\log f(x)}{\mathrm{d}x}.$$

Then, assuming R_n is well behaved and can be ignored, the formal second-order Edgeworth approximation to the distribution of T is given by:

$$\Pr[T_n \leq x] \approx F\left[x - \frac{a(x)}{\sqrt{n}} + \frac{2a(x)a'(x) + c(x)v(x) + v'(x) - 2b(x)}{2n}\right]. \tag{3.11}$$

Again, many technical assumptions on the random variables X_n, A_n, and B_n will be needed to prove the validity of the approximation. They seem to be satisfied, however, in actual applications.

For example, suppose z_n is distributed as $N_q(0, I)$ and y_n is a vector, independent of z_n, with zero mean and bounded higher moments. In many hypothesis testing problems the test statistics, under the null hypothesis, posess a stochastic expansion of the form:

$$T_n = z'z + \frac{A(z, y)}{\sqrt{n}} + \frac{B(z, y)}{n} + \frac{R}{n\sqrt{n}},$$

where A is linear in y (for given z) and $A(0, y) = 0$. (Again, the subscript n is dropped to simplify the notation.) Since F in this case is the chi-square distribution function with q degrees of freedom, $c(x) = (q - 2 - x)/2x$. Typically, $a(x) = \mathrm{E}(A \mid z'z = x) = 0$; $b(x)$ and $v(x)$ are usually homogeneous quadratic functions of x. Thus, using (3.11), we find an approximation of the form:

$$\Pr[T_n \leq x] \approx F\left[x + \frac{\beta_1 x + \beta_2 x^2}{n}\right], \tag{3.12}$$

where the β_i are functions of the moments of (z, y). Sargan (1980) gives a detailed derivation for the case where the stochastic expansion arises from a Taylor expansion of a function of moments. Rothenberg (1977, 1981b) analyzes the noncentral case where the mean of z is nonzero and F is the noncentral chi-square distribution.

To summarize, many econometric estimators and test statistics possess, after suitable standardization, stochastic expansions of the form (3.6). It is usually easy to demonstrate that R_n satisfies a regularity condition like (3.8) and the remainder term can be ignored. A formal second-order Edgeworth expansion for the truncated variable T' can be obtained from its moments, using any of the algorithms discussed above. For most econometric applications, the limiting distributions are normal or chi-square and the correction terms A_n and B_n are polynomials in asymptotically normal random variables. Thus the formal Edgeworth approximation is relatively easy to calculate, as we shall demonstrate

in later sections. Proofs that the approximation error is indeed $o(n^{-1})$ are much harder. The results of Sargan (1976, 1980) and Phillips (1977b) cover most of the cases met in practice.

4. Second-order comparisons of estimators

In any econometric inference problem, many different ways to estimate the unknown parameters are available. Since the exact sampling distributions are often unknown, choice among the alternative estimators has traditionally been based on asymptotic approximations. Typically, however, there are a number of estimators having the same limiting distributions. In those cases, second-order Edgeworth approximations can be used to distinguish among the asymptotically equivalent procedures. Indeed, a rich and powerful theory of second-order estimation efficiency has developed recently in the statistical literature. Although most of the results concern single-parameter estimation from simple random sampling, the extension of this theory to typical econometric problems is apparent.

Second-order comparisons of estimators based on moments calculated from the first few terms of stochastic expansions have been employed extensively in econometrics after the pioneering work of Nagar (1959). Some recent examples are Amemiya (1980), Fuller (1977), and Taylor (1977). Since the estimators being examined often do not possess finite moments, the status of such comparisons has been questioned by Srinivasan (1970) and others. However, if the calculated expectations are interpreted as the moments of an approximating distribution, it does not seem unreasonable to use them for comparison purposes. In fact, Pfanzagl and Wefelmeyer (1978a) show that most of the general conclusions derivable from second-order moment calculations can be restated in terms of Edgeworth approximations to the quantiles of the probability distributions.

A more serious objection to the econometric work using Nagar-type moment calculations is the lack of strong results. When the alternative estimators have different biases, mean-square-error comparisons typically are inconclusive. No estimator is uniformly best to second order. The comparisons, however, take on new meaning when interpreted in light of the general theory of second-order efficiency. This theory, although initiated over fifty years ago by R. A. Fisher (1925) and explored by C. R. Rao (1961, 1963), has reached maturity only within the past decade. The summary presented here is based on Akahira and Takeuchi (1981), Efron (1975), Ghosh and Subramanyam (1974), and Pfanzagl and Wefelmeyer (1978a, 1979).[5]

[5]Many of these statisticians use the term "second-order" to describe expansions with error $o(n^{-1/2})$ and would refer to our $o(n^{-1})$ Edgeworth approximations as "third-order". Hence, they speak of third-order efficiency.

4.1. General approach

For simplicity, we begin by considering the one-dimensional case under quadratic loss. An unknown parameter θ is to be estimated from observations on a random vector y whose joint probability distribution is $f(y, \theta)$. Under exact sample theory we would evaluate an estimator $\hat{\theta}$ by its mean square error $\mathrm{E}(\hat{\theta} - \theta)^2$. When exact distributions are unavailable, we consider a sequence of estimation problems indexed by the sample size n and use limiting distributions as approximations. For most applications, the commonly proposed estimators converge in probability to the true parameter at rate $n^{-1/2}$ and the standardized estimators are asymptotically normal. These estimators can be evaluated using expectations calculated from the approximating normal distributions. We shall denote such expectations by the symbol E_1.

Suppose $\sqrt{n}(\hat{\theta} - \theta)$ converges to a $N[\mu(\theta), \sigma^2(\theta)]$ distribution where μ and σ^2 are continuous functions of θ in a neighborhood of the true parameter value. Then, first-order mean square error $\mathrm{E}_1(\hat{\theta} - \theta)^2$ is given by $[\mu^2(\theta) + \sigma^2(\theta)]n^{-1}$. We define $\mathscr{S}_1$ to be the set of all such asymptotically normal estimators and consider the problem of finding the best estimator in $\mathscr{S}_1$. Under certain regularity conditions on the density f, the inverse information term can be shown to be a lower bound for the approximate mean square error. That is, for all $\hat{\theta}$ in $\mathscr{S}_1$:

$$n\mathrm{E}_1(\hat{\theta} - \theta)^2 \geq \frac{1}{\lambda(\theta)},$$

where

$$\lambda(\theta) = -\lim \frac{1}{n} \mathrm{E} \frac{\partial^2 \log f(y, \theta)}{\partial \theta^2}$$

is the limiting average information term for f.

An estimator in $\mathscr{S}_1$ whose approximate mean square error attains the lower bound is called asymptotically efficient. Typically, the standardized maximum likelihood estimator $\sqrt{n}(\hat{\theta}_{\mathrm{M}} - \theta)$ converges to a $N[0, \lambda^{-1}(\theta)]$ distribution and hence $\hat{\theta}_{\mathrm{M}}$ is asymptotically efficient. Of course, any other estimator which is asymptotically equivalent to $\hat{\theta}_{\mathrm{M}}$ will share this property; for example, if $\sqrt{n}(\hat{\theta}_{\mathrm{M}} - \hat{\theta})$ converges in probability to zero, then $\hat{\theta}$ and $\hat{\theta}_{\mathrm{M}}$ will be approximated by the same normal distribution and have the same first-order properties. Under suitable smoothness conditions, minimum distance estimators, Bayes estimators from arbitrary smooth priors, and linearized maximum likelihood estimators are all asymptotically efficient. [See, for example, Rothenberg (1973).] It seems natural to compare these estimators using second-order asymptotic approximations.

Let $\hat{\theta}$ be an estimator which, after standardization, possesses an asymptotic expansion of the form:

$$\sqrt{n}(\hat{\theta}-\theta) = X_n + \frac{A_n}{\sqrt{n}} + \frac{B_n}{n} + \frac{R_n}{n\sqrt{n}}, \tag{4.1}$$

where X_n, A_n, and B_n are random variables with bounded moments and limiting distributions as n tends to infinity. Suppose the limiting distribution of X_n is $N[0,\lambda^{-1}(\theta)]$ and A_n, B_n, and R_n are well behaved so that $\hat{\theta}$ is asymptotically efficient and has a distribution which can be approximated by a valid $o(n^{-1})$ Edgeworth expansion. We shall denote by $\mathscr{S}_2$ the set of all such estimators. Expectations calculated from the second-order approximate distributions will be denoted by E_2; thus, $\mathrm{E}_2(\hat{\theta}-\theta)^2$ is the mean square error when the actual distribution of $\hat{\theta}$ is replaced by the $o(n^{-1})$ Edgeworth approximation. These "second-order" moments are equivalent to those obtained by Nagar's technique of term-by-term expectation of the stochastic expansion (4.1).

4.2. *Optimality criteria*

Since the maximum likelihood estimator has minimum (first-order) mean square error in the set $\mathscr{S}_1$, it is natural to ask whether it has minimum second-order mean square error in $\mathscr{S}_2$. The answer, however, is no. If $\hat{\theta}$ is an estimator in $\mathscr{S}_2$ and θ_0 is some constant in the parameter space, then $\hat{\theta}(1-n^{-1})+\theta_0 n^{-1}$ is also in $\mathscr{S}_2$ and has lower mean square error than $\hat{\theta}$ when θ is close to θ_0. Thus, there cannot be a uniformly best estimator in $\mathscr{S}_2$ under the mean square error criterion.

Following the traditional exact theory of optimal inference [for example, Lehmann (1959, ch. 1)], two alternative approaches are available for studying estimators in $\mathscr{S}_2$. We can give up on finding a "best" estimator and simply try to characterize a minimal set of estimators which dominate all others; or we can impose an unbiasedness restriction thus limiting the class of estimators to be considered. The two approaches lead to similar conclusions.

When comparing two estimators in $\mathscr{S}_2$, it seems reasonable to say that $\hat{\theta}_1$ is as good as $\hat{\theta}_2$ if $\mathrm{E}_2(\hat{\theta}_1-\theta)^2 \leq \mathrm{E}_2(\hat{\theta}_2-\theta)^2$ for all θ. If the inequality is sometimes strict, we shall say that $\hat{\theta}_2$ is dominated by $\hat{\theta}_1$. When searching for a good estimator, we might reasonably ignore all estimators which are dominated. Furthermore, nothing is lost by excluding estimators which have the same mean square error as ones we are keeping. Suppose $\bar{\mathscr{S}}_2$ is a subset of $\mathscr{S}_2$ such that, for every estimator excluded, there is an estimator included which is as good. Since, in terms of mean square error, one cannot lose by restricting the search to $\bar{\mathscr{S}}_2$, such a set is called *essentially complete*. The characterization of (small) essentially complete classes is, according to one school of thought, the main task of a theory of estimation.

Unfortunately, essentially complete classes are typically very large and include many unreasonable estimators. If one is willing to exclude from consideration all estimators which are biased, a great simplification occurs. Although all the estimators in $\mathscr{S}_2$ are first-order unbiased, they generally are not second-order unbiased. Let $\hat{\theta}_r$ be an estimator in $\mathscr{S}_2$. Its expectation can be written as:

$$\mathrm{E}_2(\hat{\theta}_r) = \theta + \frac{b_r(\theta)}{n} + O(n^{-2}).$$

Although b_r will generally depend on the unknown parameter, it is possible to construct a second-order unbiased estimator by using the estimated bias function. Define

$$\theta_r^* = \hat{\theta}_r - \frac{b_r(\hat{\theta}_r)}{n}$$

to be the bias-adjusted estimator based on $\hat{\theta}_r$. If $b_r(\theta)$ possesses a continuous derivative, the bias-adjusted estimator has a stochastic expansion

$$\sqrt{n}\,(\theta_r^* - \theta) = \left[1 - \frac{b_r'(\theta)}{n}\right]\sqrt{n}\,(\hat{\theta}_r - \theta) - \frac{b_r(\theta)}{\sqrt{n}} + \frac{R^*}{n\sqrt{n}},$$

where b_r' is the derivative of b_r and R^* is a remainder term satisfying the regularity condition (3.8). Thus, θ_r^* is a second-order unbiased estimator in $\mathscr{S}_2$. All estimators in $\mathscr{S}_2$ with smooth bias functions can be adjusted in this way and hence we can construct the subset $\mathscr{S}_2^*$ of all second-order unbiased estimators. If unbiasedness is a compelling property, the search for a good estimator could be restricted to $\mathscr{S}_2^*$.

4.3. *Second-order efficient estimators*

In the larger class $\mathscr{S}_1$ of all (uniformly) asymptotically normal estimators, the maximum likelihood estimator $\hat{\theta}_\mathrm{M}$ is first-order minimum variance unbiased; it also, by itself, constitutes an essentially complete class of first-order minimum mean square error estimators. The extension of this result to the set $\mathscr{S}_2$ is the basis for the so-called "second-order efficiency" property of maximum likelihood estimators. Under certain regularity conditions (which take pages to state), it is possible to prove a theorem with the following conclusion:

The bias-adjusted maximum likelihood estimator

$$\theta_\mathrm{M}^* = \hat{\theta}_\mathrm{M} - \frac{b_\mathrm{M}(\hat{\theta}_\mathrm{M})}{n} \tag{4.2}$$

has smallest second-order variance among the set $\mathscr{S}_2^*$ of second-order unbiased estimators possessing $o(n^{-1})$ Edgeworth expansions. Furthermore, the class $\bar{\mathscr{S}}_M$ of all estimators of the form:

$$\hat{\theta}_M + \frac{c(\hat{\theta}_M)}{n}, \tag{4.3}$$

where c is any smooth function, is essentially second-order complete in $\mathscr{S}_2$.

For formal statements and proofs of such a theorem under i.i.d. sampling, the reader is directed to Ghosh, Sinha, and Wieand (1980) and Pfanzagl and Wefelmeyer (1978a).

This basic result of second-order estimation theory does *not* say that the maximum likelihood estimator is optimal. Indeed, it does not say anything at all about $\hat{\theta}_M$ itself. If one insists on having an unbiased estimator, then the adjusted MLE θ_M^* is best. Otherwise, the result implies that, in searching for an estimator with low mean square error (based on second-order approximations to sampling distributions), nothing is lost by restricting attention to certain functions of $\hat{\theta}_M$. The choice of an estimator from the class $\bar{\mathscr{S}}_M$ depends on one's trade-off between bias and variance; or, from a Bayesian point of view, on one's prior. Although commonly referred to as a result on second-order *efficiency*, the theorem really says no more than that $\hat{\theta}_M$ is second-order *sufficient* for the estimation problem at hand.

Of course, the second-order optimality properties of the maximum likelihood estimator are shared by many other estimators. If, for $c(\cdot)$ ranging over the set of smooth functions, the class of estimators $\hat{\theta} + c(\hat{\theta})n^{-1}$ is essentially complete in $\mathscr{S}_2$, then $\hat{\theta}$ is said to be second-order efficient. In addition to the MLE, any Bayes estimate calculated from a symmetric loss function and a smooth prior is second-order efficient. Any estimator possessing the same $o(n^{-1})$ Edgeworth expansion as an estimator in $\bar{\mathscr{S}}_M$ is also second-order efficient.

Although the set of second-order efficient estimators is large, it does not include all first-order efficient estimators. Linearized maximum likelihood and minimum distance estimators are generally dominated by functions of Bayes and ML estimators. Indeed, most common procedures used to avoid maximizing the likelihood function in nonlinear models turn out to be second-order inefficient. For example, as pointed out by Akahira and Takeuchi (1981), two-stage and three-stage least squares estimators in overidentified simultaneous equations models with normal errors are dominated by adjusted limited information and full information maximum likelihood estimators.

A full characterization of the set of second-order efficient estimators is difficult. However, it is interesting to note that, although the single iteration method of scoring (linearized maximum likelihood) does not generally lead to second-order

efficient estimators, a two-iteration scoring procedure does. Since the line of reasoning is widely used in the literature, a sketch of the argument may be worthwhile. Suppose the logarithmic likelihood function $L(\theta) \equiv \log f(y, \theta)$ possesses well-behaved derivatives so that valid stochastic expansions can be developed. In particular, for $r = 1$, 2, and 3, define $L_r = \mathrm{d}^r L(\theta)/\mathrm{d}\theta^r$; we assume L_r/n converges to the constant λ_r as n approaches infinity and that the standardized derivatives $\sqrt{n}(L_r/n - \lambda_r)$ have limiting normal distributions. Let $\hat{\theta}_0$ be some consistent estimator such that $\sqrt{n}(\hat{\theta}_0 - \hat{\theta}_M)$ has a limiting distribution. Consider the following iterative procedure for generating estimates starting from $\hat{\theta}_0$:

$$\hat{\theta}_{s+1} = \hat{\theta}_s - \frac{L_1(\hat{\theta}_s)}{L_2(\hat{\theta}_s)}, \qquad s = 0, 1, 2, \ldots.$$

Assuming an interior maximum so that $L_1(\hat{\theta}_M) = 0$, we can expand by Taylor series around $\hat{\theta}_M$ obtaining:

$$\begin{aligned}
\hat{\theta}_{s+1} - \hat{\theta}_M &= \hat{\theta}_s - \hat{\theta}_M - \frac{L_2(\hat{\theta}_M)(\hat{\theta}_s - \hat{\theta}_M) + \frac{1}{2}L_3(\hat{\theta}_M)(\hat{\theta}_s - \hat{\theta}_M)^2 + \cdots}{L_2(\hat{\theta}_M) + L_3(\hat{\theta})(\hat{\theta}_s - \hat{\theta}_M) + \cdots} \\
&= \frac{L_3(\hat{\theta}_n)(\hat{\theta}_s - \hat{\theta}_M)^2}{2L_2(\hat{\theta}_M)} + \cdots \\
&= \frac{\lambda_3(\theta)}{2\lambda_2(\theta)}(\hat{\theta}_s - \hat{\theta}_M)^2 + \cdots.
\end{aligned}$$

If $\lambda_3(\hat{\theta}_0 - \hat{\theta}_M)$ is of order $n^{-1/2}$, $\sqrt{n}(\hat{\theta}_1 - \hat{\theta}_M)$ is of order $n^{-1/2}$ and $\sqrt{n}(\hat{\theta}_2 - \hat{\theta}_M)$ is of order $n^{-3/2}$. Thus, the second iterate is second-order equivalent to $\hat{\theta}_M$ and the first iterate is not. The first iterate is second-order efficient only if $\lambda_3 = 0$ or if $\hat{\theta}_0$ is asymptotically efficient.

4.4. *Deficiency*

Second-order inefficient estimators are not necessarily poor, since the efficiency loss may be quite small. It is therefore useful to get an idea of the magnitudes involved. Hodges and Lehmann (1970) propose an interesting measure of second-order inefficiency. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two asymptotically efficient estimators in $\mathscr{S}_2$ and consider their bias-adjusted variants

$$\theta_1^* = \hat{\theta}_1 - \frac{b_1(\hat{\theta}_1)}{n} \quad \text{and} \quad \theta_2^* = \hat{\theta}_2 - \frac{b_2(\hat{\theta}_2)}{n}.$$

Suppose that θ_1^* is second-order optimal and the two bias-adjusted estimators have second-order variances of the form:

$$E_2(\theta_1^* - \theta)^2 = \frac{1}{n\lambda(\theta)}\left[1 + \frac{B_1(\theta)}{n}\right] + o(n^{-2}),$$

$$E_2(\theta_2^* - \theta)^2 = \frac{1}{n\lambda(\theta)}\left[1 + \frac{B_2(\theta)}{n}\right] + o(n^{-2}),$$

where the common asymptotic variance is the inverse information term λ^{-1} and $B_2(\theta) \geq B_1(\theta)$ for all θ. The deficiency of θ_2^* is defined to be the additional observations Δ needed, when using θ_2^*, to obtain the same precision as when using θ_1^*. That is, Δ is the solution of:

$$\frac{1}{n}\left[1 + \frac{B_1(\theta)}{n}\right] = \frac{1}{n+\Delta}\left[1 + \frac{B_2(\theta)}{n+\Delta}\right].$$

Solving, we find:

$$\Delta = B_2(\theta) - B_1(\theta) + o(n^{-1}). \tag{4.4}$$

Thus, deficiency is approximately n times the proportional difference in second-order variance.

Although deficiency is defined in terms of the bias-adjusted estimators, it can be calculated from the unadjusted estimators. If $\hat{\theta}_1$ is second-order efficient and $n(\hat{\theta}_2 - \hat{\theta}_1)$ has a limiting distribution with variance V, the deficiency of θ_2^* is $V\lambda$. The proof is based on the fact that $n(\theta_2^* - \theta_1^*)$ must be asymptotically uncorrelated with $\sqrt{n}(\theta_1^* - \theta)$; otherwise, a linear combination of θ_1^* and θ_2^* would have smaller second-order variance than θ_1^*. Hence using $o(n^{-1})$ Edgeworth approximations to compute moments:

$$\begin{aligned}\operatorname{Var}(\theta_2^*) - \operatorname{Var}(\theta_1^*) &= \operatorname{Var}(\theta_2^* - \theta_1^*) \\ &= \frac{1}{n^2}\operatorname{Var}[n(\hat{\theta}_2 - \hat{\theta}_1) + b_1(\hat{\theta}_1) - b_2(\hat{\theta}_2)] \\ &= \frac{1}{n^2}\operatorname{Var}[n(\hat{\theta}_2 - \hat{\theta}_1)] + o(n^{-2}).\end{aligned}$$

This result implies that, as far as deficiency is concerned, the key feature which distinguishes the estimators in $\mathcal{S}_2$ is the $n^{-1/2}$ term in their stochastic expansions.

4.5. *Generalizations*

Since the results on second-order efficiency presented in this section have been expressed in terms of Edgeworth expansions, they do not apply to estimation

problems for discrete probability distributions. However, the theory can be generalized to include the discrete case. Furthermore, there is no need to restrict ourselves to one-parameter problems under quadratic loss. Pfanzagl and Wefelmeyer (1978a) develop a general theory of multiparameter second-order estimation efficiency for arbitrary symmetric loss functions. When loss is a smooth function of the estimation error, both continuous and discrete distributions are covered. The conclusions are similar to the ones reported here: the set $\mathscr{S}_{\mathrm{M}}$ of maximum likelihood estimators adjusted as in (4.3) constitute an essentially complete class of estimators possessing stochastic expansions to order n^{-1}. Bayes and other estimators having the same $o(n^{-1})$ stochastic expansions share this second-order efficiency property. Although general proofs are available only for the case of simple random sampling, it is clear that the results have much broader applicability.

The fact that the second-order optimality properties of maximum likelihood and Bayes estimators hold for arbitrary symmetric loss functions is rather surprising. It suggests that there is no additional information in the third and fourth cumulants that is relevant for the second-order comparison of estimators. In fact, it has been shown by Akahira and Takeuchi (1981) that all well-behaved first-order efficient estimators necessarily have the same skewness and kurtosis to a second order of approximation. The $o(n^{-1})$ Edgeworth expansions for the estimators in $\mathscr{S}_2$ differ only by location and dispersion. Hence, for the purpose of comparing estimators on the basis of second-order approximations to their distributions, nothing is lost by concentrating on the first two moments.

Since the second-order theory of estimation is based on asymptotic expansions, the results can be relied on only to the extent that such expansions give accurate approximations to the true distributions. Clearly, if the tail behavior of estimators is really important, then the second-order comparisons discussed here are unlikely to be useful: there is no reason to believe that Edgeworth-type approximations are very accurate outside the central ninty percent of the distribution. Furthermore, in many cases where two asymptotically efficient estimators are compared, the bias difference is considerably larger than the difference in standard deviations. This suggests that correction for bias may be more important than second-order efficiency considerations when choosing among estimators.

5. Second-order comparisons of tests

5.1. General approach

The theory of second-order efficient tests of hypotheses parallels the theory for point estimation. Again, Edgeworth approximations are used in place of the traditional (first-order) asymptotic approximations to the distributions of sample

statistics. We shall consider the case where the probability law for the observed data depends on θ, a q-dimensional vector of parameters which are to be tested, and on ω, a p-dimensional vector of nuisance parameters. The null hypothesis $\theta = \theta_0$ is examined using some test statistic, T, where large values are taken as evidence against the hypothesis. The rejection region $T > t$ is said to have size α if

$$\sup_{\omega} \Pr[T > t] = \alpha,$$

when $\theta = \theta_0$; the constant t is called the critical value for the test. The quality of such a test is measured by its power: the probability that T exceeds the critical value when the null hypothesis is false.

Often the exact distribution of T is not known and the critical value is determined using the asymptotic distribution for large sample size n. Suppose, for example, that under the null hypothesis:

$$\Pr[T \leq x] = F(x) + o(n^0),$$

where the approximate distribution function F does not depend on the unknown parameters. In most applications F turns out to be the chi-square distribution with q degrees of freedom, or, when $q = 1$, simply the standard normal. The asymptotic critical value t_α for a test of size α is the solution to the equation $F(t_\alpha) = 1 - \alpha$. The test which rejects the null hypothesis when $T > t_\alpha$ is asymptotically similar of level α; that is, its type-I error probability, to a first order of approximation, equals α for all values of the nuisance parameter ω.

To approximate the power function using asymptotic methods, some normalization is necessary. The test statistic T typically has a limiting distribution when the null hypothesis is true, but not when the null hypothesis is false. For most problems, the probability distribution for T has a center which moves off to infinity at the rate $\sqrt{n}\,\|\theta - \theta_0\|$ and power approaches unity at an exponential rate as the sample size increases. Two alternative normalization schemes have been proposed in the statistical literature. One approach, developed by Bahadur (1960, 1967) and applied by Geweke (1981a, 1981b) in econometric time-series analysis, employs large deviation theory and measures, in effect, the exponential rate at which power approaches unity. The other approach, due to Pitman and developed by Hodges and Lehmann (1956) and others, considers sequences where the true parameter θ converges to θ_0 and hence examines only local or contiguous alternatives. We shall restrict our attention to this latter approach.

The purpose of the analysis is to compare competing tests on the basis of their abilities to distinguish the hypothesized value θ_0 from alternative values. Of greatest interest are alternatives in the range where power is moderate, say between 0.2 and 0.9. Outside that region, the tests are so good or so poor that

comparisons are uninteresting. Therefore, to get good approximations in the central region of the power function, it seems reasonable to treat $\sqrt{n}(\theta - \theta_0)$ as a vector of moderate values when doing the asymptotics. The local approach to approximating power functions finds the limiting distribution of the test statistic T, allowing the true parameter value θ to vary with n so that $\sqrt{n}(\theta - \theta_0)$ is always equal to the constant vector δ. Under such a sequence of local alternatives, the rejection probability approaches a limit:

$$\lim_{n \to \infty} \Pr[T > t_\alpha] = \pi_T(\delta), \tag{5.1}$$

which is called the local power function. Actual power would be approximated by $\pi_T[\sqrt{n}(\theta - \theta_0)]$. The limit (5.1) depends, of course, on θ_0, ω, and t_α in addition to δ; for notational simplicity, these arguments have been suppressed in writing the function π_T.

For many econometric inference problems, there are a number of alternative tests available, all having the same asymptotic properties. For example, as discussed by Engle in Chapter 13 of this Handbook, the Wald, Lagrange multiplier, and likelihood ratio statistics for testing multidimensional hypotheses in smooth parametric models are all asymptotically chi-square under the null hypothesis and have the same limiting noncentral chi-square distribution under sequences of local alternatives. That is, all three tests have the same asymptotic critical value t_α and the same local power functions. Yet, as Berndt and Savin (1977) and Evans and Savin (1982) point out, the small-sample behavior of the tests are sometimes quite different. It seems reasonable, therefore, to develop higher-order asymptotic expansions of the distributions and to use improved approximations when comparing tests.

Suppose the probability distribution function for T can be approximated by an Edgeworth expansion so that, under the null hypothesis:

$$\Pr[T \le t] = F\left[t + \frac{P(t,\theta,\omega)}{\sqrt{n}} + \frac{Q(t,\theta,\omega)}{n}\right] + o(n^{-1}). \tag{5.2}$$

Since F is just the (first-order) limiting distribution, the approximation depends on the unknown parameters only through the functions P and Q. Using these functions and parameter estimates $\hat{\theta}$ and $\hat{\omega}$, a modified critical value,

$$t_T^* = t_\alpha + \frac{g(t_\alpha,\hat{\theta},\hat{\omega})}{\sqrt{n}} + \frac{h(t_\alpha,\hat{\theta},\hat{\omega})}{n}, \tag{5.3}$$

can be calculated so that, when the null hypothesis is true:

$$\Pr[T > t_T^*] = \alpha + o(n^{-1}),$$

for all values of ω. The critical region $T > t_T^*$ is thus second-order similar of level α. Algorithms for determining the functions g and h from P and Q are given by Cavanagh (1983) and Pfanzagl and Wefelmeyer (1978b). The function g is simply minus P; the function h depends on the method employed for estimating the unknown parameters appearing in g. In the special (but common) case where P is zero, h is simply minus Q.

If the distribution function for T possesses a second-order Edgeworth expansion under a sequence of local alternatives where $\sqrt{n}(\theta - \theta_0)$ is fixed at the value δ as n approaches infinity, the rejection probability can be written as:

$$\Pr[T > t_T^*] \approx \pi_T(\delta) + \frac{a_T(\delta)}{\sqrt{n}} + \frac{b_T(\delta)}{n} \equiv \pi_T^*(\delta), \tag{5.4}$$

where the approximation error is $o(n^{-1})$. The function π_T^* is the second-order local power function for the size-adjusted test; again, for notational convenience, the dependency on θ_0, ω, and t_α is suppressed. By construction, $\pi_T^*(0) = \alpha$.

Suppose S and T are two alternative asymptotically equivalent test statistics possessing Edgeworth expansions. Since the two tests based on the asymptotic critical value t_α will not usually have the same size to order n^{-1}, it is not very interesting to compare their second-order power functions. It makes more sense to construct the size-adjusted critical values t_S^* and t_T^* and to compare tests with the correct size to order n^{-1}. If $\pi_T^* \geq \pi_S^*$ for all relevant values of δ and ω, then the size-adjusted test with rejection region $T > t_T^*$ is at least as good as the test with rejection region $S > t_S^*$. If the inequality is sometimes strict, then the test based on T dominates the test based on S. A second-order similar test of level α is said to be second-order efficient if it is not dominated by any other such test.

5.2. *Some results when* $q = 1$

When a single hypothesis is being tested, approximate power functions for the traditional test statistics can be written in terms of the cumulative normal distribution function. If, in addition, only one-sided alternatives are contemplated, considerable simplification occurs and a comprehensive theory of second-order optimal tests is available. The pioneering work is by Chibisov (1974) and Pfanzagl (1973); the fundamental paper by Pfanzagl and Wefelmeyer (1978b) contains the main results. More elementary expositions can be found in the survey papers by Pfanzagl (1980) and Rothenberg (1982) and in the application to nonlinear regression by Cavanagh (1981).

Suppose θ is a scalar parameter and the null hypothesis $\theta = \theta_0$ is tested against the alternative $\theta > \theta_0$. As before, ω is a vector of unknown nuisance parameters not involved in the null hypothesis. We consider test statistics whose distributions

are asymptotically normal and possess second-order Edgeworth approximations. We reject the null hypothesis for large values of the test statistic.

The Wald, Lagrange multiplier, and likelihood ratio principles lead to asymptotically normal test statistics when the distribution function for the data is well behaved. Let $L(\theta, \omega)$ be the log-likelihood function and let $L_\theta(\theta, \omega)$ be its derivative with respect to θ. Define $(\hat{\theta}, \hat{\omega})$ and $(\theta_0, \hat{\omega}_0)$ to be the unrestricted and restricted maximum likelihood estimates, respectively. If $\sqrt{n}(\hat{\theta} - \theta)$ has a limiting $N[0, \sigma^2(\theta, \omega)]$ distribution, then the Wald statistic for testing $\theta = \theta_0$ is:

$$\mathrm{W} = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sigma(\hat{\theta}, \hat{\omega})}.$$

The Lagrange multiplier statistic is:

$$\mathrm{LM} = \frac{1}{\sqrt{n}} L_\theta(\theta_0, \hat{\omega}_0)\sigma(\theta_0, \hat{\omega});$$

the likelihood ratio statistic is:

$$\mathrm{LR} = \pm\sqrt{2}\left[L(\hat{\theta}, \hat{\omega}) - L(\theta_0, \hat{\omega}_0)\right]^{1/2},$$

where the sign is taken to be the sign of W.

Under suitable smoothness assumptions, all three test statistics have limiting $N(\delta/\sigma, 1)$ distributions under sequences of local alternatives where $\theta = \theta_0 + \delta/\sqrt{n}$. Hence, if one rejects for values of the statistic greater than t_α, the upper α significance point for a standard normal, all three tests have the local power function $\Phi[\delta/\sigma - t_\alpha]$. This function can be shown to be an upper bound for the asymptotic power of any level α test. Thus the three tests are asymptotically efficient.

Using (5.3), any asymptotically efficient test of level α can be adjusted so that it is second-order similar of level α. Let $\mathscr{T}$ be the set of all such size-adjusted tests. Any test in $\mathscr{T}$ has a power function which can be expanded as in (5.4). Since all the tests in $\mathscr{T}$ are asymptotically efficient, the leading term in the power function expansion is given by $\Phi(\delta/\sigma - t_\alpha)$. It can be shown that the next term $a_T(\delta)/\sqrt{n}$ is also independent of the test statistic T. Power differences among the tests in $\mathscr{T}$ are of order n^{-1}. Furthermore, there is an upper bound on the term $b_T(\delta)/n$. For any asymptotically efficient size-adjusted test based on a statistic T, the local power function can be written as:

$$\pi_T^*(\delta) = \pi(\delta) + \frac{a(\delta)}{\sqrt{n}} + \frac{b(\delta)}{n} - \frac{c(\delta)(\delta - \delta_T)^2 + d_T(\delta)}{n} + o(n^{-1}), \qquad (5.5)$$

where $c(\delta)$ and $d_T(\delta)$ are non-negative. The functions π, a, b, and c are determined by the likelihood function and the significance level α. To second order, power depends on the test statistic T only through the constant δ_T and the function d_T.

The first three terms of (5.5) constitute the envelope power function:

$$\pi^*(\delta) = \pi(\delta) + \frac{a(\delta)}{\sqrt{n}} + \frac{b(\delta)}{n},$$

which is an upper bound to the (approximate) power of any test in $\mathscr{T}$. A test based on T is second-order efficient if, and only if, $d_T(\delta)$ is identically zero. If $c(\delta)$ is identically zero, all second-order efficient tests have the same power curve $\pi^*(\delta)$ to order n^{-1}. If $c(\delta) > 0$, second-order efficient tests have crossing power curves, each tangent to $\pi^*(\delta)$ at some point δ_T.

In any given one-tailed, one-parameter testing problem, two key questions can be asked. (i) Is $c(\delta)$ identically zero so that all second-order efficient tests have the same power function? (ii) Is the test statistic T second-order efficient and, if so, for what value δ_T is it tangent to the envelope power curve? The classification of problems and test statistics according to the answers to these questions is explored by Pfanzagl and Wefelmeyer (1978b). The value of c depends on the relationship between the first and second derivatives of the log-likelihood function. Generally, tests of mean parameters in a normal linear model have $c = 0$; tests of mean parameters in non-normal and nonlinear models have $c > 0$. Often $c(\delta)$ can be interpreted as a measure of curvature or nonlinearity. Tests based on second-order inefficient estimators have positive d and are dominated. The Lagrange multiplier, likelihood ratio, and Wald tests (based on maximum likelihood estimators) are all second-order efficient. The tangency points for the three tests are $\delta_{\mathrm{LM}} = 0$, $\delta_{\mathrm{LR}} = \sigma t_\alpha$, and $\delta_{\mathrm{W}} = 2\sigma t_\alpha$. Thus, the LM test dominates all others when power is approximately α; the Wald test dominates all others when power is approximately $1 - \alpha$; the LR test dominates at power approximately one-half.

When the alternative hypothesis is $\theta \neq \theta_0$ and $\delta = \sqrt{n}(\theta - \theta_0)$ can be negative as well as positive, the theory of optimal tests is more complicated. If the test statistic T is asymptotically distributed as $N(\delta/\sigma, 1)$, it is natural to reject when T assumes either large positive or large negative values. Using the Edgeworth approximation to the distribution of T, one can find critical values t_1 and t_2 [generally functions of the data as in (5.3)] such that the test which rejects when T lies outside the interval $(-t_1, t_2)$ is second-order similar of level α; that is, when $\theta = \theta_0$:

$$\Pr[-t_1 \leq T \leq t_2] = 1 - \alpha + o(n^{-1}), \tag{5.6}$$

for all ω. Indeed, for any statistic T, there are infinitely many pairs (t_1, t_2) satisfying (5.6), each yielding a different power curve. If, for example, t_2 is considerably greater than t_1, power will be high for negative δ and low for positive δ. Unless some restrictions are placed on the choice of rejection region, no uniformly optimal test is possible.

If the null distribution of T were symmetric about zero, the symmetric rejection region with $t_1 = t_2$ would be a natural choice. However, since the Edgeworth approximation to the distribution of T is generally skewed, the symmetric region is not particularly compelling. Sargan (1975b), in the context of constructing approximate confidence intervals for structural coefficients in the simultaneous equations model, suggested minimizing the expected length $t_1 + t_2$ of the acceptance region. Unfortunately, since the t_i depend on the unknown parameters, he concludes that this criterion is nonoperational and, in the end, recommends the symmetric region.

An alternative approach is to impose the restriction of unbiasedness. This is the basis of much traditional nonasymptotic testing theory [see, for example, Lehmann (1959, ch. 4)] and is commonly employed in estimation theory. A test is said to be unbiased if the probability of rejecting when the null hypothesis is false is always at least as large as the probability of rejecting when the null hypothesis is true. That is, the power function takes its minimum value when $\theta = \theta_0$. In the case $q = 1$, the condition that the test be locally unbiased uniquely determines t_1 and t_2 to second order. Size-adjusted locally unbiased Wald, likelihood ratio, and Lagrange multiplier tests are easily constructed from the Edgeworth expansions. If the curvature measure $c(\delta)$ is zero, the three tests have identical power functions to order n^{-1}; if $c(\delta) \neq 0$, the power functions cross. Again, as in the one-tailed case, the Wald test dominates when power is high and the LR test dominates when power is near one-half. However, the LM test is no longer second-order efficient; when $c(\delta) \neq 0$, one can construct locally unbiased tests having uniformly higher power. Details are given in Cavanagh and Rothenberg (1983).

5.3. *Results for the multiparameter case*

There does not exist a comprehensive theory of second-order optimality of tests when $q > 1$. However, numerous results are available for special cases. Peers (1981), Hayakawa (1975), and others have analyzed multidimensional null hypotheses for arbitrary smooth likelihood functions. These studies have concentrated on power differences of order $n^{-1/2}$. Fujikoshi (1973), Ito (1956, 1960), and Rothenberg (1977, 1981b) have investigated the normal linear model using expansions to order n^{-1}. In this latter case, after adjustment for size, tests based on the Wald, likelihood ratio, and Lagrange multiplier principles (using maximum

likelihood estimates) are all second-order efficient. However, their power surfaces generally cross and no one test is uniformly best.

The major findings from these studies do not concern differences in power, but differences in size. When q is large, the adjusted critical values (5.3) based on the Edgeworth approximation often differ dramatically from the asymptotic values. The null distributions of typical test statistics for multidimensional hypotheses are not at all well approximated by a chi-square. This phenomenon has been noted by Evans and Savin (1982), Laitinen (1978), Meisner (1979), and others.

Consider, for example, the multivariate regression model $Y = X\Pi + V$, where Y is an $n \times q$ matrix of observations on q endogenous variables, X is an $n \times K$ matrix of observations on K exogenous variables, and Π is a $K \times q$ matrix of regression coefficients. The n rows of V are i.i.d. vectors distributed as $N_q(0, \Omega)$. Suppose the null hypothesis is that $\Pi' a = 0$, for some K-dimensional vector a. The Wald test statistic is:

$$T = \frac{a'\hat{\Pi}\hat{\Omega}^{-1}\hat{\Pi}'a}{a'(X'X)^{-1}a},$$

where $\hat{\Pi} = (X'X)^{-1}X'Y$ and $\hat{\Omega} = Y'[I - X(X'X)^{-1}X']Y/(n-K)$. Define $m = n - K$. Then it is known that, for $m \geq q$, T is a multiple of an F-statistic. In fact, $T(m-q+1)/mq$ is distributed as F with q and $m-q+1$ degrees of freedom. The usual asymptotic approximation to the distribution of T is a chi-square with q degrees of freedom. The mean of T is actually $mq/(m-q-1)$, whereas the mean of the chi-square is q; even if m is five times q, the error is more than twenty percent. Clearly, very large samples will be needed before the asymptotic approximation is reasonable when many restrictions are being tested.

For most testing problems, the exact distributions and the errors in the asymptotic approximation are unknown. However, it is reasonable to assume that higher-order asymptotic expansions will lead to improved approximations. The second-order Edgeworth approximation typically takes the simple chi-square form (3.12). In multivariate normal linear models (with quite general covariance structures allowing for autocorrelation and heteroscedasticity), the size-adjusted critical values (5.3) for the Wald, Lagrange multiplier, and likelihood ratio tests on the regression coefficients can be written as:

$$t_{\mathrm{W}}^{*} = t_{\alpha}\left[1 + \frac{\gamma_1 + \gamma_0 t_{\alpha}}{n}\right],$$

$$t_{\mathrm{LM}}^{*} = t_{\alpha}\left[1 + \frac{\gamma_2 - \gamma_0 t_{\alpha}}{n}\right],$$

$$t_{\mathrm{LR}}^{*} = t_{\alpha}\left[1 + \frac{\gamma_1 + \gamma_2}{2n}\right],$$

where t_α is the upper α quantile for a chi-square distribution with q degrees of freedom. The coefficients γ_0, γ_1, and γ_2 depend on the particular problem; some examples are given in Rothenberg (1977, 1981b). For this normal regression case, the likelihood ratio has the simplest correction since, to second order, the test statistic is just a multiple of a chi-square. It also turns out to be (approximately) the arithmetic average of the other two statistics. These adjusted critical values appear to be reasonably accurate when q/n is small (say, less than 0.1). Even when q/n is larger, they seem to be an improvement on the asymptotic critical values obtained from the chi-square approximation.

Although the second-order theory of hypothesis testing is not yet fully developed, work completed so far suggests the following conclusions. For moderate sample sizes, the actual significance levels of commonly used tests often differ substantially from the nominal level based on first-order asymptotic approximations. Modified critical regions, calculated from the first few terms of an Edgeworth series expansion of the distribution functions, are available and have significance levels much closer to the nominal level. Unmodified Wald, likelihood ratio, and Lagrange multiplier test statistics, although asymptotically equivalent, often assume very different numerical values. The modified (size-adjusted) tests are less likely to give conflicting results. The formulae for the modified critical regions are relatively simple (for likelihood-ratio tests it typically involves a constant degrees-of-freedom adjustment); their use should be encouraged, especially when the number of restrictions being tested is large.

Once the tests have been modified so that they have the same significance level (to a second order of approximation), it is possible to compare their (approximate) power functions. Here the differences often seem to be small and may possibly be swamped by the approximation error. However, when substantial differences do appear, the second-order theory provides a basis for choice among alternative tests. For example, in nonlinear problems, it seems that the likelihood-ratio test has optimal power characteristics in the interesting central region of the power surface and may therefore be preferable to the Lagrange multiplier and Wald tests.

5.4. *Confidence intervals*

The critical region for a test can serve as the basis for a confidence region for the unknown parameters. Suppose $T(\theta_0)$ is a test statistic for the q-dimensional null hypothesis $\theta = \theta_0$ against the alternative $\theta \neq \theta_0$. Let t^* be a size-adjusted critical value such that, when the null hypothesis is true,

$$\Pr\left[T(\theta_0) > t^*\right] = \alpha + o(n^{-1}),$$

for all values of the nuisance parameters. If $T(\theta_0)$ is defined for all θ_0 in the parameter space, we can form the set $C = \{\theta' : T(\theta') < t^*\}$ of all hypothesized values that are not rejected by the family of tests based on T. By construction the random set C covers the true parameter with probability $1-\alpha+o(n^{-1})$ and hence is a valid confidence region at that level.

A good confidence region covers incorrect parameter values with low probability. Thus, good confidence regions are likely to result from powerful tests. If the test is locally unbiased, then so will be the confidence region: it covers the true θ with higher probability than any false value nearby.

When $q=1$, the results described in Section 5.2 can be applied to construct locally unbiased confidence regions for a scalar parameter. For example, one might use the locally unbiased Lagrange multiplier or likelihood ratio critical region to define a confidence set. Unfortunately, the sets of θ_0 values satisfying

$$-t_1 \le \mathrm{LM}(\theta_0) \le t_2 \quad \text{or} \quad -t_1 \le \mathrm{LR}(\theta_0) \le t_2,$$

are sometimes difficult to determine and need not be intervals. The Wald test, however, always leads to confidence intervals of the form

$$C_{\mathrm{W}} = \left\{\theta : \hat{\theta} - \frac{t_2}{\sqrt{n}}\sigma(\hat{\theta},\hat{\omega}) \le \theta \le \hat{\theta} + \frac{t_1}{\sqrt{n}}\sigma(\hat{\theta},\hat{\omega})\right\},$$

where t_1 and t_2 are determined by the requirement that the test be locally unbiased and second-order similar of level α. These critical values take the form:

$$t_1 = t + \frac{p(t,\hat{\theta},\hat{\omega})}{\sqrt{n}} + \frac{q(t,\hat{\theta},\hat{\omega})}{n},$$

$$t_2 = t - \frac{p(t,\hat{\theta},\hat{\omega})}{\sqrt{n}} + \frac{q(t,\hat{\theta},\hat{\omega})}{n},$$

where t is the asymptotic critical value satisfying $1-\Phi(t)=\alpha/2$ and the functions p and q depend on the Edgeworth expansion of the test statistic. Since C_{W} is easy to construct, it would be the natural choice in practice.

6. Some examples

6.1. *Simultaneous equations*

Consider the two-equation model:

$$y_1 = \alpha y_2 + u; \qquad y_2 = z + v, \tag{6.1}$$

where α is an unknown scalar parameter, y_1 and y_2 are n-dimensional vectors of

observations on two endogenous variables, and u and v are n-dimensional vectors of unobserved errors. The vector $z = \mathrm{E}y_2$ is unknown, but is assumed to lie in the column space of the observed nonrandom $n \times K$ matrix Z having rank K. The n pairs (u_i, v_i) are independent draws from a bivariate normal distribution with zero means, variances σ_u^2 and σ_v^2, and correlation coefficient ρ.

The first equation in (6.1) represents some structural relationship of interest; the second equation is part of the reduced form and, in the spirit of limited information analysis, is not further specified. Additional exogenous explanatory variables could be introduced in the structural equation without complicating the distribution theory; additional endogenous variables require more work. We shall discuss here only the simple case (6.1). Exact and approximate distribution theory for general simultaneous equations models is covered in detail by Phillips in Chapter 8 of Volume I of this Handbook.[6] Our purpose is merely to illustrate some of the results given in Sections 3–5 above.

The two-stage least squares (2SLS) estimator $\hat{\alpha}$ is defined as:

$$\hat{\alpha} = \frac{y_2' N y_1}{y_2' N y_2} = \alpha + \frac{z'u + v'Nu}{z'z + 2z'v + v'Nv},$$

where $N = Z(Z'Z)^{-1}Z'$ is the rank-K idempotent projection matrix for the column space of Z. It will simplify matters if $\hat{\alpha}$ is expressed in terms of a few standardized random variables:

$$X = \frac{z'u}{\sigma_u \sqrt{z'z}}; \qquad Y = \frac{z'v}{\sigma_v \sqrt{z'z}},$$

$$s = \frac{v'Nu}{\sigma_v \sigma_u}; \qquad S = \frac{v'Nv}{\sigma_v^2}.$$

The pair (X, Y) is bivariate normal with zero means, unit variances, and correlation coefficient ρ. The random variable s has mean $K\rho$ and variance $K(1+\rho^2)$; S has mean K and variance $2K$. The standardized two-stage least squares estimator is:

$$d \equiv \frac{\sqrt{z'z}}{\sigma_u}(\hat{\alpha} - \alpha) = \frac{X + (s/\mu)}{1 + (2Y/\mu) + (S/\mu^2)} \tag{6.2}$$

[6] To simplify the analysis, Phillips (and others) transform the coordinate system and study a canonical model where the reduced-form errors are independent. Since the original parameterization is retained here, our formulae must be transformed to agree with theirs.

where $\mu^2 = z'z/\sigma_v^2$ is often called the "concentration" parameter. When μ is large, d behaves like the $N(0,1)$ random variable X. Note that the sample size n affects the distribution of d only through the concentration parameter. For asymptotic analysis it will be convenient to index the sequence of problems by μ rather than the traditional $n^{1/2}$. Although large values of μ would typically be due to a large sample size, other explanations are possible; e.g. a very small value for σ_v^2. Thus, large μ asymptotics can be interpreted as either large n or small σ asymptotics.

Since the two-stage least squares estimator in our model is an elementary function of normal random variables, exact analysis of its distribution is possible. An infinite series representation of its density function is available; see, for example, Sawa (1969). Simple approximations to the density and distribution functions are also easy to obtain. If μ is large (and K is small), the denominator in the representation (6.2) should be close to one with high probability. This suggests developing the stochastic power-series expansion:

$$d = X + \frac{s - 2XY}{\mu} + \frac{4XY^2 - XS - 2Ys}{\mu^2} + \frac{R}{\mu^3}, \tag{6.3}$$

and using the first three terms (denoted by d') to form a second-order Edgeworth approximation. Since R satisfies the regularity condition (3.8), the algorithms given in Section 3 (with μ replacing $n^{1/2}$) are available. To order $o(\mu^{-2})$, the first two moments of d' are:

$$\mathrm{E}(d') = \frac{(K-2)\rho}{\mu}, \qquad \mathrm{Var}(d') = 1 - \frac{(K-4)(1+3\rho^2)+4\rho^2}{\mu^2};$$

the third and fourth cumulants are approximately

$$k_3 = -\frac{6\rho}{\mu}; \qquad k_4 = \frac{12(1+5\rho^2)}{\mu^2}.$$

Substitution into (3.7) yields the Edgeworth-B approximation:

$$\Pr\left[\frac{\sqrt{z'z}}{\sigma_u}(\hat{\alpha} - \alpha) \le x\right] \approx \Phi\left[x + \frac{\rho(x^2+1-K)}{\mu} + \frac{x(K-1)(1-\rho^2) + x^3(3\rho^2-1)}{2\mu^2}\right]. \tag{6.4}$$

This result was obtained by Sargan and Mikhail (1971) and by Anderson and Sawa (1973) using somewhat different methods.

Calculations by Anderson and Sawa (1979) indicate that the Edgeworth approximation is excellent when μ^2 is greater than 50 and K is 3 or less.[7] The approximation seems to be adequate for any μ^2 greater than 10 as long as K/μ is less than unity. When K exceeds μ, however, the approximation often breaks down disastrously. Of course, it is not surprising that problems arise when K/μ is large. The terms s/μ and S/μ^2 are taken to be small compared to X in the stochastic expansion (6.3). When K is the same order as μ, this is untenable and an alternative treatment is required.

The simplest approach is to take advantage of the ratio form of d. Equation (6.2) implies:

$$\Pr[d \leq x] = \Pr\left[X + \frac{s - 2xY}{\mu} - \frac{xS}{\mu^2} \leq x\right]$$
$$\equiv \Pr[W(x) \leq x].$$

Since N is idempotent with rank K, s and S behave like the sum of K independent random variables. When K is large, a central limit theorem can be invoked to justify treating them as approximately normal. Thus, for any value of x and μ, W is the sum of normal and almost normal random variables. The mean $m(x)$ and variance $\sigma^2(x)$ of W can be calculated exactly. Treating W as normal yields the approximation:

$$\Pr[d \leq x] \approx \Phi\left[\frac{x - m(x)}{\sigma(x)}\right].$$

A better approximation could be obtained by calculating the third and fourth cumulants of $(W - m)/\sigma$ and using the Edgeworth approximation to the distribution of W in place of the normal. Some trial calculations indicate that high accuracy is attained when K is 10 or more. Unlike the Edgeworth approximation (6.4) applied directly to d', Edgeworth applied to W improves with increasing K.

Many other methods for approximating the distribution of d are available. Holly and Phillips (1979) derive the saddle-point approximation and conclude that it performs better than (6.4), particularly when K is large. Phillips (1981) experiments with fitting rational functions of the form (2.1) and reports excellent results. In both cases, the density function is approximated analytically and numerical integration is required for the approximation of probabilities. The Edgeworth approximation (6.4) generalizes easily to the case where there are many endogenous explanatory variables; indeed, the present example is based on

[7]Anderson and Sawa actually evaluate the Edgeworth-A form of the approximation, but the general conclusions presumably carry over to (6.4).

the paper by Sargan and Mikhail (1971) which covers the general case. Alternative approximation methods, like the one exploiting the ratio form of d, do not generalize easily. Likewise, Edgeworth approximations can be developed for estimators which do not have simple representations, situations where most other approximating methods are not applicable.

Numerous estimators for α have been proposed as alternatives to two-stage least squares. The second-order theory of estimation described in Section 4 can be employed to compare the competing procedures. Consider, for example, Theil's k-class of estimators:

$$\hat{\alpha}_k = \frac{y_2'(I - kM)y_1}{y_2'(I - kM)y_2},$$

where $M = I - N$. The maximum likelihood estimator is a member of this class when $k_{\mathrm{ML}} - 1$ is given by λ, the smallest root of the determinental equation:

$$\left|(y_1, y_2)'(N - \lambda M)(y_1, y_2)\right| = 0.$$

This root possesses a stochastic expansion of the form:

$$\lambda = \frac{u'(N - N_z)u + A_1\mu^{-1} + A_2\mu^{-2} + \cdots}{u'Mu},$$

where $N_z = z(z'z)^{-1}z'$ and the A_i are stochastically bounded as μ tends to infinity.

The standardized k-class estimator can be written as:

$$d_k \equiv \frac{\sqrt{z'z}}{\sigma_u}(\hat{\alpha}_k - \alpha) = \frac{X + (s_k/\mu)}{1 + (2Y/\mu) + (S_k/\mu^2)},$$

where

$$s_k = s + (1 - k)\frac{u'Mv}{\sigma_u\sigma_v},$$

$$S_k = S + (1 - k)\frac{v'Mv}{\sigma_v^2}.$$

If s_k and S_k are (stochastically) small compared to μ, d_k can be expanded in a power series analogous to (6.3) and its distribution approximated by a second-order Edgeworth expansion. Since $u'Mv$ and $v'Mv$ have means and variances proportional to $n - K$, such expansions are reasonable only if $(n - K)(k - 1)$ is

small. We shall examine the special class where

$$k = 1 + a\lambda - \frac{b}{n-K}, \tag{6.5}$$

for constants a and b of moderate size (compared to μ). When $a = b = 0$, the estimator is 2SLS; when $a = 1$ and $b = 0$, the estimator is LIML. Thus our subset of k-class estimators includes most of the interesting cases.

The truncated power-series expansion of the standardized k-class estimator is given by:

$$d_k' = X + \frac{s_k - 2XY}{\mu} + \frac{4XY^2 + XS_k - 2Ys_k}{\mu^2};$$

its first two moments, to order μ^{-2}, are:

$$\mathrm{E}(d_k') = \rho\frac{(1-a)(K-1)+b-1}{\mu},$$

$$\mathrm{Var}(d_k') = 1 - \frac{(K-4)(1+3\rho^2)+4\rho^2+2b(1+2\rho^2)-2a(1+a\rho^2)(K-1)}{\mu^2},$$

as long as the sample size n is large compared to the number K.[8] To order $o(\mu^{-2})$, the third and fourth cumulants of d_k' are the same as given above for 2SLS. This implies that all k-class estimators of the form (6.5) have the same skewness and kurtosis to second order. The $o(\mu^{-2})$ Edgeworth approximations differ only with respect to location and dispersion, as implied by the general result stated in Section 4.5.

The mean square error for the standardized estimator d_k' is the sum of the variance and the squared mean. The expression is rather messy but depends on a and b in a very systematic way. Indeed, we have the following striking result first noted by Fuller (1977): the family of k-class estimators with $a = 1$ and $b \geq 4$ is essentially complete to second order. Any other k-class estimator is dominated by a member of that family.

The k-class estimator is unbiased, to second order, if and only if $b = 1 + (a-1)(K-1)$. In that case, the variance becomes:

$$1 + \frac{K(1-\rho^2)+2\rho^2+2\rho^2(K-1)(a-1)^2}{\mu^2}, \tag{6.6}$$

[8]When K/n is not negligible, an additional term is needed in the variance formula; the relative merits of the maximum likelihood estimator are reduced. Cf. Anderson (1977) and Morimune (1981). Sargan (1975a) and Kunitomo (1982) have developed an asymptotic theory for "large models" where K tends to infinity along with the sample size n.

which clearly is minimized when $a=1$. The estimator with $a=b=1$ is therefore best unbiased; it is second-order equivalent to the bias-adjusted MLE. The estimator with $a=0$, $b=2-K$ is also approximately unbiased; it is second-order equivalent to the bias-adjusted 2SLS estimator. From (6.6), the deficiency of the bias-adjusted 2SLS estimator compared to the bias-adjusted MLE is:

$$\Delta = n\frac{2(K-1)\rho^2}{\mu^2} = 2(K-1)\rho^2\frac{1-r^2}{r^2},$$

where $r^2 = z'z/(z'z+n\sigma_v^2)$ is the population coefficient of determination for the reduced-form equation. The number of observations needed to compensate for using the second-order inefficient estimator is greatest when K is large and the reduced-form fit is poor; that is, when Z contains many irrelevant or collinear variables.

These results on the relative merits of alternative k-class estimators generalize to the case where the structural equation contains many explanatory variables, both endogenous and exogenous. If y_2 is replaced by the matrix Y_2 and α is interpreted as a vector of unknown parameters, the k-class structural estimator is:

$$\hat{\alpha}_k = \left[Y_2'(I-kM)Y_2\right]^{-1}Y_2'(I-kM)y_1.$$

Let one vector estimator be judged better than another if the difference in their mean square error matrices is negative semidefinite. Then, with λ interpreted as the root of the appropriate determinantal equation, the family of k-class estimators (6.5) with $a=1$ and $b\geq 4$ is again essentially complete to second order; any other k-class estimator is dominated. The estimator with $a=b=1$ is again best second-order unbiased. These results can be established by the same methods employed above, with matrix power-series expansions replacing the scalar expansions.

The second-order theory of hypothesis testing can be applied to the simultaneous equations model. The calculation of size-adjusted critical regions and second-order power functions is lengthy and will not be given here. However, the following results can be stated. Tests based on 2SLS estimates are dominated by likelihood based tests as long as $K>1$. The curvature measure $c(\delta)$ defined in Section 5 is zero for the problem of testing $\alpha=\alpha_0$ in the model (6.1). Hence, Wald, LM, and LR tests, after size correction, are asymptotically equivalent. In more complex, full-information models the curvature measure is nonzero and the power functions for the three tests cross. Detailed results on second-order comparisons of tests in simultaneous equations models are not yet available in print. Some preliminary findings are reported in Cavanagh (1981) and Turkington (1977). Edgeworth approximations to the distribution functions of some test statistics under the null hypothesis are given in Sargan (1975b, 1980).

6.2. *Autoregressive models*

Suppose the $n+1$ random variables $u_0, u_1, \ldots, u_n$ are distributed normally with zero means and with second moments:

$$\mathrm{E}(u_i u_j) = \sigma^2 \rho^{|i-j|}, \qquad -1 < \rho < 1.$$

The u_i are thus $n+1$ consecutive observations from a stationary first-order autoregressive process. If $u = (u_1, \ldots, u_n)'$ and $u_{-1} = (u_0, \ldots, u_{n-1})'$ are defined as n-dimensional column vectors, the model can be written in regression form:

$$u = \rho u_{-1} + \varepsilon,$$

where ε is $N(0, \sigma_\varepsilon^2 I)$, with $\sigma_\varepsilon^2 = \sigma^2(1-\rho^2)$. The least squares regression coefficient $u'u_{-1}/u'_{-1}u_{-1}$ is often used to estimate ρ. Approximations to its sampling distribution are developed by Phillips (1977a, 1978). We shall consider here the modified estimator:

$$\hat{\rho} = \frac{u'u_{-1}}{u'_{-1}u_{-1} + (u_n^2 - u_0^2)/2}, \tag{6.7}$$

which treats the end points symmetrically. It has the attractive property of always taking values in the interval $(-1, 1)$. Since $\sqrt{n}(\hat{\rho} - \rho)$ has a limiting normal distribution with variance $1-\rho^2$, it is natural to analyze the standardized statistic which can be written as:

$$T = \frac{\sqrt{n}(\hat{\rho} - \rho)}{\sqrt{1-\rho^2}} = \left(X - \frac{\alpha Z}{\sqrt{n}}\right)\left(1 + \frac{Y}{\sqrt{n}}\right)^{-1},$$

where $\alpha = \rho/\sqrt{1-\rho^2}$ and

$$X = \frac{u'_{-1}\varepsilon}{\sigma\sigma_\varepsilon\sqrt{n}}; \qquad Y = \frac{u'_{-1}u_{-1} + (u_n^2 - u_0^2)/2 - n\sigma^2}{\sigma^2\sqrt{n}}; \qquad Z = \frac{u_n^2 - u_0^2}{2\sigma^2}.$$

Even though X and Y are not sums of independent random variables, they have limiting normal distributions and possess rth order cumulants of order $n^{1-r/2}$ for $r \geq 2$. It seems reasonable to expect that the distribution of T can be approximated by a valid Edgeworth expansion using the techniques developed in Section 3. The truncated power series expansion for T can be written as:

$$T' = X - \frac{XY + \alpha Z}{\sqrt{n}} + \frac{XY^2 + \alpha YZ}{n}.$$

The cumulants of T' are very complicated functions of ρ, but can be approximated to $o(n^{-1})$ as long as ρ is not too close to one. Very tedious calculation yields the approximate first four cumulants:

$$\mathrm{E}(T') = \frac{-2\alpha}{\sqrt{n}}; \qquad \mathrm{Var}(T') = 1 + \frac{7\alpha^2 - 2}{n},$$

$$k_3 = \frac{-6\alpha}{\sqrt{n}}; \qquad k_4 = \frac{6(10\alpha^2 - 1)}{n}.$$

From the general formula (3.7), the Edgeworth-B approximation to the distribution function is:

$$\Pr[T \leq t] \approx \Phi\left[t + \frac{\alpha(t^2+1)}{\sqrt{n}} + \frac{t(1+4\alpha^2) + t^3(1+6\alpha^2)}{4n}\right]. \tag{6.8}$$

The Edgeworth approximation is based on the idea that the high-order cumulants of the standardized statistic are small. When $\alpha = 1$ (that is, $\rho = 0.7$), the third cumulant is approximately $-6/\sqrt{n}$ and the fourth cumulant is approximately $54/n$. A high degree of accuracy cannot be expected for sample sizes less than, say, 50. Numerical calculations by Phillips (1977a) and Sheehan (1981) indicate that the Edgeworth approximation is not very satisfactory in small samples when ρ is greater than one half.

The poor performance of the Edgeworth approximation when ρ is large stems from two sources. First, when autocorrelation is high the distribution of $\hat{\rho}$ is very skewed and does not look at all like a normal. Second, the approximations to the cumulants are not accurate when ρ is near one since they drop "end point" terms of the form ρ^n. The former difficulty can be alleviated by considering normalizing transformations which reduce skewness.

A transformation which performs this task and is interesting for its own sake is:

$$T^* = \frac{\frac{n+1}{n}\hat{\rho} - \rho}{\sqrt{(1-\hat{\rho}^2)/(n-1)}}.$$

The numerator is the difference between the bias adjusted estimator and the true parameter value; the denominator is the estimated standard error. The ratio can be interpreted as a modified Wald statistic under the null hypothesis. Since T^* has the power series expansion

$$T^* \approx T + \frac{\alpha(1+T^2)}{\sqrt{n}} + \frac{(1+2\alpha^2)T + (1+3\alpha^2)T^3}{2n},$$

its distribution can be approximated by reverting the series and using (6.8). The Edgeworth-B approximation to the distribution function is given by:

$$\Pr[T^* \le x] \approx \Pr\left[T \le x - \frac{\alpha(1+x^2)}{\sqrt{n}} - \frac{(1-2\alpha^2)x + (1-\alpha^2)x^3}{2n}\right]$$
$$\approx \Phi\left[x - \frac{x+x^3}{4n}\right],$$

which is the Edgeworth approximation to a Student-t distribution. When ρ is large, T^* is not centered at zero and a more complicated bias adjustment is needed. Otherwise, treating T^* as a Student-t random variable with n degrees of freedom seems to give a reasonable approximation for moderate sample sizes.

Many other approaches are available for approximating the distributions of sample statistics from autoregressive models. Anderson (1971, ch. 6) surveys some of the statistical literature. Alternative inverse trigonometric transformations are considered by Jenkins (1954), Quenouille (1948), and Sheehan (1981). Saddle-point methods are employed by Daniels (1956), Durbin (1980), and Phillips (1978). Various curve fitting techniques are also possible. For small values of ρ, all the methods seem reasonably satisfactory. When ρ is large, the general purpose methods seem to break down; ad hoc approaches which take into account the special features of the problem then are necessary.

A slightly more complicated situation occurs when the u_i are unobserved regression errors. Consider the linear model $y = X\beta + u$, where the n-dimensional error vector is autocorrelated as before; the $n \times k$ matrix of regressors are assumed nonrandom. Let $\tilde{\beta} = (X'X)^{-1}X'y$ be the least squares estimator of β and let $\hat{\beta}$ be the maximum likelihood estimator.[9] The two n-dimensional residual vectors are defined as:

$$\tilde{u} = y - X\tilde{\beta} \quad \text{and} \quad \hat{u} = y - X\hat{\beta};$$

their lagged values are:

$$\tilde{u}_{-1} = y_{-1} - X_{-1}\tilde{\beta} \quad \text{and} \quad \hat{u}_{-1} = y_{-1} - X_{-1}\hat{\beta},$$

where y_{-1} and X_{-1} are defined in the obvious way. We assume observations on the variables are available for period 0, but are not used in estimating β; this makes the notation easier and will not affect the results.

Two alternative estimators for ρ are suggested. The residuals $\tilde{u}$ could be used in place of u in (6.7); or the residuals $\hat{u}$ could be used. For purposes of comparison,

[9]Any estimator asymptotically equivalent to the MLE will have the same properties. For example, $\hat{\beta}$ might be the generalized least squares estimator based on some consistent estimator for ρ.

the end-point modifications play no role so we shall consider the two estimators

$$\tilde{\rho} = \frac{\tilde{u}'\tilde{u}_{-1}}{\tilde{u}'_{-1}\tilde{u}_{-1}} \quad \text{and} \quad \hat{\rho} = \frac{\hat{u}'\hat{u}_{-1}}{\hat{u}'_{-1}\hat{u}_{-1}}.$$

Both are asymptotically efficient, but $\hat{\rho}$ dominates to second order since it is based on an efficient estimator of β. Sheehan (1981) calculates the deficiency of the bias-adjusted estimator based on the least-squares residuals. He shows that the deficiency is equal to the asymptotic variance of $n(\tilde{\rho} - \hat{\rho})$ divided by the asymptotic variance of $\sqrt{n}(\hat{\rho} - \rho)$.

The computation of these variances is lengthy, but the basic method can be briefly sketched. By definition, $X'\tilde{u} = 0$ and $\tilde{u} = \hat{u} - X(\tilde{\beta} - \hat{\beta})$. Furthermore, the maximum likelihood normal equations imply that $(X - \hat{\rho}X_{-1})'(u - \hat{\rho}u_{-1}) \approx 0$. After considerable manipulation of these equations, the standardized difference in estimators can be written, to a first order of approximation, as:

$$\begin{aligned} n(\tilde{\rho} - \hat{\rho}) &= \frac{n}{\tilde{u}'_{-1}\tilde{u}_{-1}}\left[\tilde{u}'\tilde{u}_{-1} - \hat{u}'\hat{u}_{-1} + \hat{\rho}\left(\hat{u}'_{-1}\hat{u}_{-1} - \tilde{u}'_{-1}\tilde{u}_{-1}\right)\right] \\ &\approx \frac{1}{\rho\sigma^2}(\tilde{\beta} - \hat{\beta})'\left(\rho X'X_{-1} - X'X\right)(\tilde{\beta} - \hat{\beta}). \end{aligned}$$

Defining the $K \times K$ matrices:

$$A = X'(X - \rho X_{-1}),$$

$$V = \mathrm{E}(\tilde{\beta} - \hat{\beta})(\tilde{\beta} - \hat{\beta})' \approx (X'X)^{-1}X'\Sigma X(X'X)^{-1} - (X'\Sigma^{-1}X)^{-1},$$

where $\Sigma = \mathrm{E}(uu')$, deficiency has the simple expression:

$$\Delta = \frac{\operatorname{tr} AVAV + \operatorname{tr} A'VAV}{\rho^2(1-\rho^2)\sigma^4}. \tag{6.9}$$

In the special case where β is a scalar and the single regressor vector x is itself an autoregressive process with autocorrelation coefficient r, the deficiency formula simplifies to:

$$\Delta = \frac{8\rho^2}{1-\rho^2}\left[1 + \frac{(r-\rho)^2}{1-r^2}\right]^{-2},$$

which is bounded by $8\rho^2/(1-\rho^2)$. For example, when both u and x have autocorrelation coefficients near 0.7, the use of least squares residuals in place of maximum likelihood residuals is equivalent to throwing away eight observations.

6.3. Generalized least squares

Consider the linear regression model:

$$y = X\beta + u, \tag{6.10}$$

where X is an $n \times K$ matrix of nonrandom regressors and u is an n-dimensional vector of normal random errors with zero mean. The error covariance matrix is written as $E(uu') = \Omega^{-1}$, where the precision matrix Ω depends on the vector ω of p unknown parameters. For a given estimator $\hat{\omega}$, the estimated precision matrix is $\hat{\Omega} = \Omega(\hat{\omega})$. The generalized least squares (GLS) estimator based on the estimate $\hat{\omega}$ is:

$$\hat{\beta} = (X'\hat{\Omega}X)^{-1}X'\hat{\Omega}y. \tag{6.11}$$

Under suitable regularity conditions on X and Ω, the standardized estimator $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal and possesses a valid second-order Edgeworth expansion as long as $\sqrt{n}(\hat{\omega} - \omega)$ has a limiting distribution. Furthermore, since the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta)$ does not depend on which estimator $\hat{\omega}$ is used, all such GLS estimators are first-order efficient. It therefore follows from the general proposition of Akahira and Takeuchi that they all must have the same third and fourth cumulants up to $o(n^{-1})$. Since the estimator using the true ω is exactly normal, we have the surprising result: to a second order of approximation, all GLS estimators based on well-behaved estimates of ω are normally distributed.

It is possible to develop $o(n^{-2})$ expansions for generalized least squares estimators. Suppose $\hat{\omega}$ is an even function of the basic error vector u and has a probability distribution not depending on the parameter β. (The maximum likelihood estimator of ω and all common estimators based on least squares residuals have these properties.) Let c be an arbitrary K-dimensional constant vector. Then, if $c'\hat{\beta}$ has variance σ^2:

$$\Pr\left[\frac{c'(\hat{\beta} - \beta)}{\sigma} \leq x\right] = \Phi\left[x - \frac{x^3 - 3x}{24n^2}a\right] + o(n^{-2}), \tag{6.12}$$

where a/n^2 is the fourth cumulant of $c'(\hat{\beta} - \beta)/\sigma$. The assumption that $\hat{\beta}$ possesses finite moments is not necessary as long as $\hat{\beta}$ has a stochastic expansion with a well-behaved remainder term; σ^2 and a then are the moments of the truncated expansion. The simplicity of the approximation (6.12) results from the following fact: If $\bar{\beta}$ is the GLS estimator using the true Ω and $\hat{\omega}$ satisfies the above-mentioned conditions, then $\hat{\beta} - \bar{\beta}$ is distributed independently of $\bar{\beta}$, is symmetric around zero, and is of order n^{-1}. Details and proofs are given in Rothenberg (1981a).

The variance σ^2 can be approximated to second order using Nagar's technique. Taylor (1977) examines a special case of the GLS model where the errors are independent but heteroscedastic. Phillips (1977c) investigates the seemingly unrelated regression model where Ω has a Kronecker product form. In this latter case, a very simple deficiency result can be stated. Suppose there are G regression equations of the form:

$$y_i = X_i\beta_i + u_i, \qquad i = 1,\ldots,G,$$

where each regression has m observations; y_i and u_i are thus m-dimensional vectors and X_i is an $m \times k_i$ matrix of nonrandom regressors. For each observation, the G errors are distributed as $N_G(0, \Sigma)$; the m error vectors for the different observations are mutually independent. The G equations can be written as one giant system of the form (6.10) where $n = Gm$ and $K = \Sigma k_i$. One might wish to compare the GLS estimator $\bar{\beta}$ (which could be used if Σ were known) with the GLS estimator $\hat{\beta}$ based on some asymptotically efficient estimate $\hat{\Sigma}$. (A common choice for $\hat{\sigma}_{ij}$ would be $\hat{u}_i'\hat{u}_j/n$ where $\hat{u}_i$ is the residual vector from an OLS regression of y_i on X_i.) Rothenberg (1981a) shows that the deficiency of $c'\hat{\beta}$ compared to $c'\bar{\beta}$ is bounded by $G+1$ and equals $G-1$ in the special case where the X_i are mutually orthogonal. Although the number of unknown nuisance parameters grows with the square of G, the deficiency grows only linearly.

6.4. *Departures from normality*

All of the above examples concern sampling from normal populations. Indeed, a search of the econometric literature reveals no application of higher-order asymptotic expansions that dispenses with the assumption of normal errors. This is rather odd, since the original intention of Edgeworth in developing his series was to be able to represent non-normal populations.

In principle, there is no reason why second-order approximations need be confined to normal sampling schemes. Although discrete lattice distributions cause some difficulties, valid Edgeworth expansions can be developed for statistics from any continuous population distribution possessing sufficient moments. The basic Edgeworth-B approximation formula (3.7) does not assume normality of the original observations. The normality assumption enters only when the approximate cumulants are computed using Nagar's technique.

In univariate problems, there seems to be no practical difficulty in dropping the normality assumption. The cumulants of the truncated statistic T' will, of course, depend on the higher cumulants of the population error distribution, but the Edgeworth approximation should be computable. In fact, one could conduct interesting studies in robustness by seeing how the approximate distribution of an estimator varies as the error distribution departs from normality.

In multivariate problems, however, things become more complex. The cumulants of T' will depend on all the third and fourth cross cumulants of the errors. Although the calculations can be made, the resulting approximation formula will be very difficult to interpret unless these cross cumulants depend in a simple way on a few parameters. The assumption that the errors are normal can be relaxed if a convenient multivariate distribution can be found to replace it.

7. Conclusions

Approximate distribution theory, like exact distribution theory, derives results from assumptions on the stochastic process generating the data. The quality of the approximation will not be better than the quality of the specifications on which it is based. The models used by econometricians are, at best, crude and rather arbitrary. One would surely not want to rely on a distribution theory unless the conclusions were fairly robust to small changes in the basic assumptions. Since most of the approximation methods discussed here employ information on the first four moments of the data whereas the usual asymptotic theory typically requires information only on the first two moments, some loss in robustness must be expected. However, if a rough idea about the degree of skewness and kurtosis is available, that information often can be exploited to obtain considerably improved approximations to sample statistics.

Clearly, sophisticated approximation theory is most appropriate in situations where the econometrician is able to make correct and detailed assumptions about the process being studied. But the theory may still be quite useful in other contexts. In current practice, applied econometricians occasionally draw incorrect conclusions on the basis of alleged asymptotic properties of their procedures. Even if the specification of the model is incomplete, second-order theory can sometimes prevent such mistakes. For example, in the presence of correlation between regressors and errors in a linear model, the two-stage least squares estimator will be strongly biased if the number of instruments is large and the instruments explain little variation in the regressors. The bias formula derived in section 6.1 under the assumption of normality may be somewhat off if the errors are in fact non-normal. But the general conclusion based on the second-order theory is surely more useful than the assertion that the estimator is consistent and hence the observed estimate should be believed.

In recent years there has developed among econometricians an extraordinary fondness for asymptotic theory. Considerable effort is devoted to showing that some new estimator or test is asymptotically normal and efficient. Of course, asymptotic theory is important in getting some idea of the sampling properties of a statistical procedure. Unfortunately, much bad statistical practice has resulted from confusing the words "asymptotic" and "approximate". The assertion that a

standardized estimator is asymptotically normal is a purely mathematical proposition about the limit of a sequence of probability measures under a set of specified assumptions. The assertion that a given estimator is approximately normal suggests that, for the particular problem at hand, the speaker believes that it would be sensible to treat the estimator as though it were really normal. Obviously, neither assertion implies the other.

Accurate and convenient approximations for the distributions of econometric estimators and test statistics are of great value. Sometimes, under certain circumstances, asymptotic arguments lead to good approximations. Often they do not. The same is true of second-order expansions based on Edgeworth or saddlepoint methods. A careful econometrician, armed with a little statistical theory, a modest computer, and a lot of common sense, can always find reasonable approximations for a given inference problem. This survey has touched on some of the statistical theory. The computer and the common sense must be sought elsewhere.

References

Akahira, M. and K. Takeuchi (1981) *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*. New York: Springer-Verlag.

Albers, W. (1978) "Testing the Mean of a Normal Population under Dependence", *Annals of Statistics*, 6, 1337–1344.

Amemiya, T. (1980) "The n^{-2}-Order Mean Squared Errors of the Maximum Likelihood and the Minimum Logit Chi-Square Estimator", *Annals of Statistics*, 8, 488–505.

Anderson, T. W. (1971) *The Statistical Analysis of Time Series*. New York: Wiley.

Anderson, T. W. (1974) "An Asymptotic Expansion of the Distribution of the Limited Information Maximum Likelihood Estimate of a Coefficient in a Simultaneous Equation System", *Journal of the American Statistical Association*, 69, 565–573.

Anderson, T. W. (1977) "Asymptotic Expansions of the Distributions of Estimates in Simultaneous Equations for Alternative Parameter Sequences", *Econometrica*, 45, 509–518.

Anderson, T. W. and T. Sawa (1973) "Distributions of Estimates of Coefficients of a Single Equation in a Simultaneous System and Their Asymptotic Expansions", *Econometrica*, 41, 683–714.

Anderson, T. W. and T. Sawa (1979) "Evaluation of the Distribution Function of the Two-Stage Least Squares Estimate", *Econometrica*, 47, 163–182.

Bahadur, R. R. (1960) "Stochastic Comparison of Tests", *Annals of Mathematical Statistics*, 31, 276–295.

Bahadur, R. R. (1967) "Rates of Convergence of Estimates and Test Statistics", *Annals of Mathematical Statistics*, 38, 303–324.

Barndorff-Nielsen, O. and D. R. Cox (1979) "Edgeworth and Saddle-Point Approximations with Statistical Applications", *Journal of the Royal Statistical Society*, B, 41, 279–312.

Berndt, E. and N. E. Savin (1977) "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model", *Econometrica*, 45, 1263–1277.

Bhattacharya, R. N. and J. K. Ghosh (1978) "On the Validity of the Formal Edgeworth Expansion", *Annals of Statistics*, 6, 434–451.

Bhattacharya, R. N. and R. R. Rao (1976) *Normal Approximations and Asymptotic Expansions*. New York: Wiley.

Bickel, P. J. (1974) "Edgeworth Expansions in Nonparametric Statistics", *Annals of Statistics*, 2, 1–20.

Bickel, P. J. and K. A. Doksum (1977) *Mathematical Statistics*. San Francisco: Holden-Day.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Cavanagh, C. L. (1981) "Hypothesis Testing in Nonlinear Models", Working Paper, University of California, Berkeley.
Cavanagh, C. L. (1983) "Hypothesis Testing in Models with Discrete Dependent Variables", Ph.D. thesis, University of California, Berkeley.
Cavanagh, C. L. and T. J. Rothenberg (1983) "The Second-Order Inefficiency of the Efficient Score Test", Working Paper, Institute of Business and Economic Research, University of California, Berkeley.
Chibisov, D. M. (1974) "Asymptotic Expansions for Some Asymptotically Optimal Tests", in: J. Hajek, ed., *Proceedings of the Prague Symposium on Asymptotic Statistics*, vol. 2, 37–68.
Chibisov, D. M. (1980) "An Asymptotic Expansion for the Distribution of a Statistic Admitting a Stochastic Expansion", I, *Theory of Probability and its Applications*, 25, 732–744.
Cornish, E. A. and R. A. Fisher (1937) "Moments and Cumulants in the Specification of Distributions", *Review of the International Statistical Institute*, 5, 307–320.
Cramer, H. (1946) *Mathematical Methods of Statistics*, Princeton: Princeton University Press.
Daniels, H. E. (1956) "The Approximate Distribution of Serial Correlation Coefficients", *Biometrika*, 43, 169–185.
Durbin, J. (1979) "Discussion of the Paper by Barndorff-Nielsen and Cox", *Journal of the Royal Statistical Society*, B, 41, 301–302.
Durbin, J. (1980) "The Approximate Distribution of Serial Correlation Coefficients Calculated from Residuals on Fourier Series", *Biometrika*, 67, 335–350.
Durbin, J. and G. Watson (1971) "Serial Correlation in Least Squares Regression", III, *Biometrika*, 58, 1–19.
Edgeworth, F. Y. (1917) "On the Mathematical Representation of Statistical Data", *Journal of the Royal Statistical Society*, 80, 411–437.
Efron, B. (1975) "Defining the Curvature of a Statistical Problem (with Applications to Second-Order Efficiency)", *Annals of Statistics*, 3, 1189–1242.
Evans, G. B. A. and N. E. Savin (1982) "Conflict Among the Criteria Revisited", *Econometrica*, forthcoming.
Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, vol. 2. New York: Wiley.
Fisher, R. A. (1925) "Theory of Statistical Estimation", *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
Fujikoshi, Y. (1973) "Asymptotic Formulas for the Distributions of Three Statistics for Multivariate Linear Hypotheses", *Annals of the Institute of Statistical Mathematics*, 25, 423–437.
Fuller, W. A. (1977) "Some Properties of a Modification of the Limited Information Estimator", *Econometrica*, 45, 939–953.
Geweke, J. (1981a) "A Comparison of Tests of the Independence of Two Covariance-Stationary Time Series", *Journal of the American Statistical Association*, 76, 363–373.
Geweke, J. (1981b) "The Approximate Slopes of Econometric Tests", *Econometrica*, 49, 1427–1442.
Ghosh, J. K., B. K. Sinha, and H. S. Wieand (1980) "Second Order Efficiency of the MLE with Respect to any Bowl-Shaped Loss Function", *Annals of Statistics*, 8, 506–521.
Ghosh, J. K. and K. Subramanyam (1974) "Second Order Efficiency of Maximum Likelihood Estimators", *Sankhya*, A, 36, 325–358.
Hayakawa, T. (1975) "The Likelihood Ratio Criteria for a Composite Hypothesis Under a Local Alternative", *Biometrika*, 62, 451–460.
Henshaw, R. C. (1966) "Testing Single-Equation Least-Squares Regression Models for Autocorrelated Disturbances", *Econometrica*, 34, 646–660.
Hodges, J. L. and E. L. Lehmann (1956) "The Efficiency of Some Nonparametric Competitors of the t-Test", *Annals of Mathematical Statistics*, 27, 324–335.
Hodges, J. L. and E. L. Lehmann (1967) Moments of Chi and Powers of t, in: L. LeCam and J. Neyman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press.
Hodges, J. L. and E. L. Lehmann (1970) "Deficiency", *Annals of Mathematical Statistics*, 41, 783–801.
Holly, A. and P. C. B. Phillips (1979) "A Saddlepoint Approximation to the Distribution of the k-Class Estimator of a Coefficient in a Simultaneous System", *Econometrica*, 47, 1527–1547.
Imhof, J. P. (1961) "Computing the Distribution of Quadratic Forms in Normal Variables", *Biometrika*, 48, 419–426

Ito, K. (1956) "Asymptotic Formulae for the Distribution of Hotelling's Generalized T^2 Statistic", *Annals of Mathematical Statistics*, 27, 1091–1105.
Ito, K. (1960) "Asymptotic Formulae for the Distribution of Hotelling's Generalized T^2 Statistic", II, *Annals of Mathematical Statistics*, 31, 1148–1153.
Jenkins, G. M. (1954) "An Angular Transformation for the Serial Correlation Coefficient", *Biometrika*, 41, 261–265.
Johnson, N. L. (1949) "Systems of Frequency Curves Generated by Methods of Translation", *Biometrika*, 36, 149–176.
Johnson, N. L. and S. Kotz (1970) *Continuous Univariate Distributions*, vol. 1. New York: Wiley.
Kadane, J. (1971) "Comparison of k-Class Estimators when Disturbances are Small", *Econometrica*, 39, 723–739.
Kendall, M. G. and A. Stuart (1969) *The Advanced Theory of Statistics*, vol. 1. London: Griffin.
Koerts, J. and A. P. J. Abrahamse (1969) *On the Theory and Application of the General Linear Model*. Rotterdam: University Press.
Kunitomo, N. (1982) Asymptotic Efficiency and Higher Order Efficiency of the Limited Information Maximum Likelihood Estimator in Large Econometric Models, Technical report 365, Institute for Mathematical Studies in the Social Sciences, Stanford University.
Laitinen, K. (1978) "Why is Demand Homogeneity so Often Rejected"? *Economics Letters*, 1, 187–191.
Lehmann, E. L. (1959) *Testing Statistical Hypotheses*. New York: Wiley.
Meisner, J. F. (1979) "The Sad Fate of the Asymptotic Slutsky Symmetry Test for Large Systems", *Economics Letters*, 2, 231–233.
Morimune, K. (1981) "Asymptotic Expansions of the Distribution of an Improved Limited Information Maximum Likelihood Estimator", *Journal of the American Statistical Association*, 76, 476–478.
Nagar, A. L. (1959) "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations", *Econometrica*, 27, 573–595.
Olver, F. W. J. (1974) *Asymptotics and Special Functions*. New York: Academic Press.
Pan Jie-jian (1968) "Distributions of the Noncircular Serial Correlation Coefficients", *Selected Translations in Mathematical Statistics and Probability*, 7, 281–292.
Peers, H. W. (1971) "Likelihood Ratio and Associated Test Criteria", *Biometrika*, 58, 577–587.
Pfanzagl, J. (1973) "Asymptotically Optimum Estimation and Test Procedures", in: J. Hajek, ed., *Proceedings of the Prague Symposium on Asymptotic Statistics*, vol. 2, 201–272.
Pfanzagl, J. (1980) "Asymptotic Expansions in Parametric Statistical Theory", in: P. R. Krishnaiah, ed., *Developments in Statistics*, vol. 3. New York: Academic Press.
Pfanzagl, J. and W. Wefelmeyer (1978a) "A third-Order Optimum Property of the Maximum Likelihood Estimator", *Journal of Multivariate Analysis*, 8, 1–29.
Pfanzagl, J. and W. Wefelmeyer (1978b) "An Asymptotically Complete Class of Tests", *Zeitschrift fur Wahrscheinlichkeitstheorie*, 45, 49–72.
Pfanzagl, J. and W. Wefelmeyer (1979) Addendum to: "A Third-Order Optimum Property of the Maximum Likelihood Estimator", *Journal of Multivariate Analysis*, 9, 179–182.
Phillips, P. C. B. (1977a) "Approximations to Some Finite Sample Distributions Associated with a First-Order Stochastic Difference Equation", *Econometrica*, 45, 463–485; erratum, 50: 274.
Phillips, P. C. B. (1977b) "A General Theorem in the Theory of Asymptotic Expansions as Approximations to the Finite Sample Distributions of Econometric Estimators", *Econometrica*, 45, 1517–1534.
Phillips, P. C. B. (1977c) "An Approximation to the Finite Sample Distribution of Zellner's Seemingly Unrelated Regression Estimator", *Journal of Econometrics*, 6, 147–164.
Phillips, P. C. B. (1978) "Edgeworth and Saddlepoint Approximations in the First-Order Noncircular Autoregression", *Biometrika*, 65, 91–98.
Phillips, P. C. B. (1980) "Finite Sample Theory and the Distributions of Alternative Estimators of the Marginal Propersity to Consume", *Review of Economic Studies*, 47, 183–224.
Phillips, P. C. B. (1981) A New Approach to Small Sample Theory, Cowles Foundation Discussion Paper, Yale University.
Quenouille, M. H. (1948) "Some Results in the Testing of Serial Correlation Coefficients", *Biometrika*, 35, 261–267.
Rao, C. R. (1961) "Asymptotic Efficiency and Limiting Information", in: *Proceedings of the Fourth*

Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. Berkeley: University of California Press.
Rao, C. R. (1963) "Criteria of Estimation in Large Samples", *Sankhya*, A, 25, 189–206.
Rothenberg, T. J. (1973) *Efficient Estimation with a Priori Information*. New Haven: Yale University Press.
Rothenberg, T. J. (1977) Edgeworth Expansions for Some Test Statistics in Multivariate Regression, Working Paper, University of California, Berkeley.
Rothenberg, T. J. (1981a) Approximate Normality of Generalized Least Squares Estimates, Working Paper, University of California, Berkeley.
Rothenberg, T. J. (1981b) Hypothesis Testing in Linear Models when the Error Covariance Matrix is Nonscalar, Working Paper, University of California, Berkeley.
Rothenberg, T. J. (1982) "Comparing Alternative Asymptotically Equivalent Tests", in: W. Hildenbrand, ed., *Advances in Econometrics*. Cambridge: Cambridge University Press.
Sargan, J. D. (1974) "On the Validity of Nagar's Expansion for the Moments of Econometric Estimators", *Econometrica*, 42, 169–176.
Sargan, J. D. (1975a) "Asymptotic Theory and Large Models", *International Economic Review*, 16, 75–91.
Sargan, J. D. (1975b) "Gram-Charlier Approximations Applied to t Ratios of k-Class Estimators", *Econometrica*, 43, 327–346.
Sargan, J. D. (1976) "Econometric Estimators and the Edgeworth Approximation", *Econometrica*, 44, 421–448; erratum, 45, 272.
Sargan, J. D. (1980) "Some Approximations to the Distribution of Econometric Criteria which are Asymptotically Distributed as Chi-Squared", *Econometrica*, 48, 1107–1138.
Sargan, J. D. and W. M. Mikhail (1971) "A General Approximation to the Distribution of Instrumental Variables Estimates", *Econometrica*, 39, 131–169.
Sawa, T. (1969) "The Exact Sampling Distributions of Ordinary Least Squares and Two Stage Least Squares Estimates", *Journal of the American Statistical Association*, 64, 923–980.
Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
Sheehan, D. (1981) Approximating the Distributions of some Time-Series Estimators and Test Statistics, Ph.D. thesis, University of California, Berkeley.
Srinivasan, T. N. (1970) "Approximations to Finite Sample Moments of Estimators Whose Exact Sampling Distributions are Unknown", *Econometrica*, 38, 533–541.
Taylor, W. E. (1977) "Small Sample Properties of a Class of Two Stage Aitken Estimators", *Econometrica*, 45, 497–508.
Theil, H. (1971) *Principles of Econometrics*. New York: Wiley.
Theil, H. and A. L. Nagar (1961) "Testing the Independence of Regression Disturbances", *Journal of the American Statistical Association*, 56, 793–806.
Turkington, D. (1977) Hypothesis Testing in Simultaneous Equations Models, Ph.D. thesis, University of California, Berkeley.
Wallace, D. L. (1958) "Asymptotic Approximations to Distributions", *Annals of Mathematical Statistics*, 29, 635–654.
Wilson, E. B. and M. M. Hilferty (1931) "The Distribution of Chi Square", *Proceedings of the National Academy of Science*, U.S.A., 17, 684–688.