DISTRIBUTION APPROXIMATION, ROTH'S THEOREM, AND LOOKING FOR INSECTS IN SHIPPING CONTAINERS

GEOFFREY DECROUEZ — UNIVERSITY OF MELBOURNE

PETER HALL — U MELBOURNE & UC DAVIS

Background

When undertaking normal approximation to the distribution of a statistic such as the estimator of a binomial proportion, where the sampling distribution is lattice, the lack of smoothness of the true distribution often has to be taken into account. This difficulty motivates the continuity correction, and also the fiducial approach taken by Clopper and Pearson (1934) and Sterne (1954) to estimating a binomial proportion.

There is also a very large, more recent literature discussing methodology for solving problems such as constructing confidence intervals for the difference or sum of two binomial proportions. Important contributions to that literature include the work of Duffy and Santer (1987), Lee et al. (1997), Agresti and Caffo (2000), Brown et al. (2001, 2002), Zhou et al. (2001), Price and Bonnett (2004), Brown and Li (2005), Borkowf (2006), Roths and Tebbs (2006), Wang (2010) and Zieliński (2010).

Cases where sample sizes are open to choice

In problems where we wish to construct a confidence interval for, or test for, the difference between two binomial proportions, the existing literature is founded on the premiss that the sample sizes are somehow chosen in advance. In such cases the error due to discontinuity is generally of the same size as the error arising from issues such as skewness, kurtosis etc.

However, if the sample sizes can be chosen by the experimenter, then there is considerable scope for selecting them so that the error due to discontinuity is of second order, even for one-sided procedures.

This talk was motivated by a problem arising when assessing the risk that environmental contaminants, in particular foreign species of insects, cross a frontier. In that setting it was desired to construct a confidence interval for the sum of two unknown binomial proportions. The experimenters were able to choose the respective sample sizes, within reasonable limits. We shall show that, in this setting, substantial reductions in the coverage error of a confidence interval, or the level inaccuracy of a hypothesis test, are possible.

THEORETICAL BACKGROUND

Let $\hat{\theta}$ denote an estimator of an unknown θ, and assume that $T = n^{1/2}\,(\hat{\theta} - \theta)/\sigma$ is asymptotically normal $\mathrm{N}(0, 1)$, where n denotes sample size. Under standard assumptions,

$$P(T \le x) = \Phi(x) + n^{-1/2}\,P(x)\,\phi(x) + o\big(n^{-1/2}\big)\,, \tag{1}$$

uniformly in x, where Φ and ϕ are the standard normal distribution and density functions, respectively, and P is an even, quadratic polynomial.

For example, if $\hat{\theta}$ denotes the mean of a sample of size n from a population with mean θ, and if data from the population have finite third moment and a nonlattice distribution, then (1) holds with

$$P(x) = \tfrac{1}{6}\,\beta\,\big(1 - x^2\big)\,,$$

where β is the standardised skewness of the population.

On the other hand, still in the case of the mean for a population with finite third moment, if the population is lattice then an extra, discontinuous term has to be added to (1).

Theoretical background – (2)

The extra term reflects the discrete continuity correction that statisticians are obliged to use when approximating a lattice distribution, for example a binomial or Poisson distribution, by the smooth normal distribution:

$$P(T \le x) = \Phi(x) + n^{-1/2}\, P(x)\, \phi(x) + n^{-1/2}\, d_n(x)\, \phi(x) + o\big(n^{-1/2}\big)\,.$$

Here $d_n(x) = (e_0/\sigma)\, \psi_n(x)$ denotes the discontinuous term in the Edgeworth expansion, e_0 is the maximal span of the lattice, σ^2 is the population variance,

$$\psi_n(x) = \psi\big\{(x - \xi_n)\, \sigma\, n^{1/2}/e_0\big\}\,, \quad \xi_n = \big(e_0/\sigma\, n^{1/2}\big)\, \big\{\tfrac{1}{2} - \psi(n\, x_0/e_0)\big\}\,,$$

$\psi(x) = \lfloor x \rfloor - x + \frac{1}{2}$, $\lfloor x \rfloor$ is the largest integer not strictly exceeding x, and it is assumed that all points of support in the distribution (of which θ is the mean) have the form $x_0 + \nu\, e_0$ for an integer ν.

Model for multi-sample cases

Our results and methodology are available only for multi-sample problems, which we now introduce.

Let X_{ji}, for $1 \le i \le n_j$ and $j = 1, \ldots, k$, denote independent random variables. Assume that each X_{ji} has a nondegenerate lattice distribution, depending on j but not on i and with maximal lattice edge width e_j and finite third moment. Suppose too that $k \ge 2$.

Put $\bar{X}_j = n_j^{-1} \sum_i X_{ji}$, $\mu_j = E(X_{ji})$, $\sigma_j^2 = \mathrm{var}(X_{ji})$ and

$$S = \sum_{j=1}^{k} \bar{X}_j \,. \tag{2}$$

The model (2) includes cases of apparently greater generality, for example where signed weights are incorporated in the series, since the absolute values of the weights can be incorporated into (2) by modifying the lattice edge widths, and negative signs can be addressed by reflecting the summand distributions.

In particular, in the case $k = 2$ the model includes the cases of a sum or difference of estimators of binomial proportions.

Properties of the model

Since third moments are finite then, if the distributions of $X_{11}, \ldots, X_{k1}$ were to satisfy a smoothness condition, such as that of Cramér, we could express the distribution of S in a one-term Edgeworth expansion:

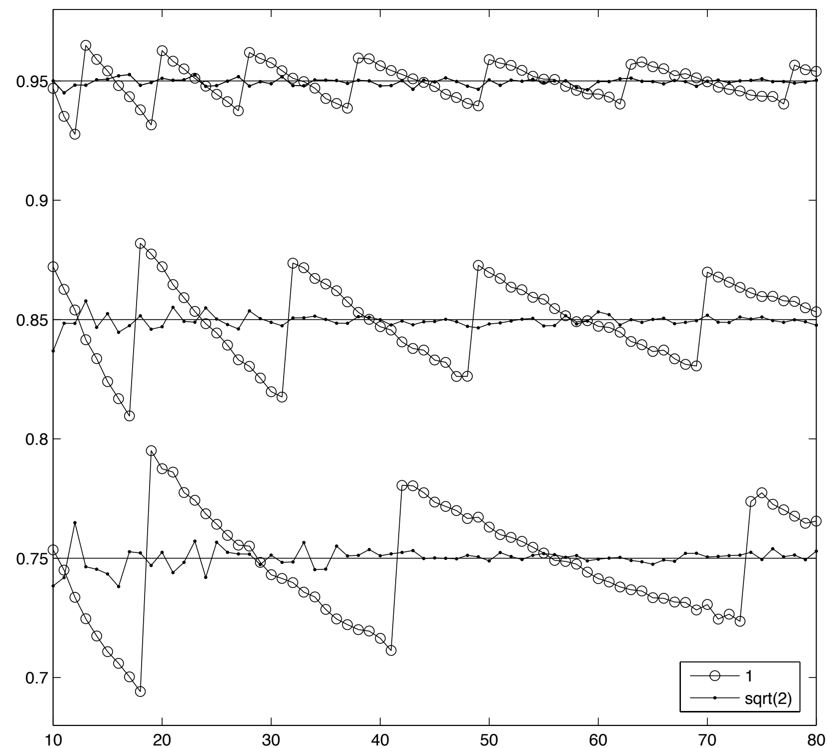
$$P\left\{\frac{S - E(S)}{(\mathrm{var} S)^{1/2}} \leq x\right\} = \Phi(x) + n^{-1/2}\,\tfrac{1}{6}\,\beta\,\big(1 - x^2\big)\,\phi(x) + o\big(n^{-1/2}\big)\,, \tag{2}$$

Here we take $n = n_1 + \ldots + n_k$ to be the asymptotic parameter, and

$$\beta = \beta(n) = \frac{n^{1/2}\,E(S - ES)^3}{(\mathrm{var}\,S)^{3/2}} = \frac{n^{1/2}\,\sum_j n_j^{-2}\,E(X_{j1} - EX_{j1})^3}{(\sum_j n_j^{-1}\,\mathrm{var}\,X_{j1})^{3/2}}$$

is a measure of standardised skewness. Under our assumptions it is bounded as $n \to \infty$.

However, in general (2) does not hold in the lattice-valued case. One way to see this is to consider the case $k = 2$ and $n_1 = n_2$. Here the need to include an extra, discontinuous term, of size $n^{-1/2}$, persists even in the asymptotic limit.



0.95
0.9
0.85
0.8
0.75
0.7
10
20
30
40
50
60
70
80
1
sqrt(2)

PROPERTIES OF THE MODEL – (2)

We assume that each $E|X_{j1}|^3 < \infty$; that each X_{j1} is distributed on a lattice $x_j + \nu\, e_j$, for integers ν, where e_j is the maximal lattice edge width; and that the sample size ratios n_{j_1}/n_{j_2} are all bounded away from zero and infinity as $n \to \infty$.

Recall the following formula, from the previous slide:

$$P\left\{\frac{S - E(S)}{(\mathrm{var} S)^{1/2}} \le x\right\} = \Phi(x) + n^{-1/2}\, \tfrac{1}{6}\, \beta\, \big(1 - x^2\big)\, \phi(x) + o\big(n^{-1/2}\big)\,. \tag{2}$$

In the first part of the theorem below we impose the following condition on at least one of the ratios $\rho_{j_1 j_2} = (e_{j_2} n_{j_1})/(e_{j_1} n_{j_2})$: As $n \to \infty$,

$$\text{for each integer } \ell \ge 1\,, \quad n^{1/2}\, |\sin(\ell \rho_{j_1 j_2} \pi)| \to \infty\,. \tag{3}$$

Theorem. (i) If, for some pair j_1, j_2 with $1 \le j_1 < j_2 \le k$, $\rho_{j_1 j_2}$ satisfies (3), then the one-term Edgeworth expansion at (2) holds uniformly in x. (ii) However, if $\rho_{j_1 j_2}$ equals a fixed rational number (not depending on n) for each pair j_1, j_2, then the expansion at (2) fails to hold because it does not include an appropriate discontinuous term of size $n^{-1/2}$.

Condition (3)

Recall condition (3):

$$\text{for each integer } \ell \geq 1\,, \quad n^{1/2}\,|\sin(\ell\rho_{j_1j_2}\pi)| \to \infty\,. \tag{3}$$

Condition (3) holds if $\rho_{j_1j_2}$ converges to an irrational number, ρ_0 say, as $n \to \infty$, since if ρ_0 is irrational then $|\sin(\ell\rho_0\pi)| > 0$ for all integers ℓ.

Condition (3) also holds if $\rho_{j_1j_2}$ converges sufficiently slowly to a rational number, for example if $\rho_{j_1j_2} = \rho_0 + \epsilon_{j_1j_2}$, where ρ_0 is rational and $\epsilon_{j_1j_2} = \epsilon_{j_1j_2}(n)$, which can be either positive or negative, converges to zero strictly more slowly than $n^{1/2}$:

$$\epsilon_{j_1j_2} \to 0\,, \quad n^{1/2}\,|\epsilon_{j_1j_2}| \to \infty\,.$$

On the other hand, if $\rho_{j_1j_2}$ converges sufficiently quickly to a rational number then not only does (3) fail, but the expansion (2) does not hold. Indeed, part (ii) of the theorem asserts that the theorem fails if $\rho_{j_1j_2}$ is a fixed rational number, for each pair (j_1, j_2) and each n.

Condition (3) also holds in many cases where $\rho_{j_1j_2}$ does not converge.

“TYPES” OF IRRATIONAL NUMBERS

The theorem shows that, if at least one of the ratios $\rho_{j_1 j_2}$ converges to an irrational number as n diverges, then the discontinuous term, of order $n^{-1/2}$, is actually of smaller order than $n^{-1/2}$.

To obtain a more concise bound on the discontinuous term we shall investigate in detail cases where one or more of the ratios $\rho_{j_1 j_2}$ converge to an irrational number as n diverges. This seems to require discussion of the “type” of an irrational.

If x is a real number, let $\langle x \rangle$ denote the distance from x to the nearest integer. In particular, if $\lfloor x \rfloor$ is the integer part function, $\langle x \rangle = \min\{x - \lfloor x \rfloor, 1 - (x - \lfloor x \rfloor)\}$.

We say that the irrational number ρ is of type η if η equals the supremum of all ζ such that $\liminf_{p \to \infty} p^{\zeta} \langle p\rho \rangle = 0$.

Properties of *convergents* of irrational numbers (convergents are particularly accurate rational approximations, based on continued fractions) can be used to prove that the type of any given irrational number always satisfies $\eta \geq 1$.

“Types” of irrational numbers – (2)

Roth’s Theorem (Roth, 1955) implies that all algebraic irrationals (that is, all irrational numbers that are roots of polynomials with rational coefficients) are of minimal type, i.e. $\eta = 1$, which is one of the cases we consider below.

More generally, if a number is chosen randomly, for example as the value of a random variable having a continuous distribution on the real line, then with probability 1 it is an irrational of type 1.

Irrationals that are not algebraic are said to be transcendental, and can have type strictly greater than 1. (However, the transcendental number e is of type 1.)

Known upper bounds to the types of the transcendental numbers π, π^2 and $\log 2$ are 6.61, 4.45 and 2.58, respectively. Liouville numbers have type $\eta = \infty$.

In our problem, irrational numbers of type 1 seem to produce fewer problems from discontinuities — so in practice (in view of Roth’s Theorem) we favour the algebraic irrationals, and e.

Refinement of theorem about effect of discontinuities

Recall that $S = \sum_j \bar{X}_j$, and that $\rho_{j_1 j_2} = (e_{j_2} n_{j_1})/(e_{j_1} n_{j_2})$. The theorem below gives conditions under which the net contribution of the discontinuous term equals $O(n^{\delta-(1/2)-(1/2\eta)})$, for all $\delta > 0$, when some $\rho_{j_1 j_2}$ is sufficiently close to an irrational number of type η.

Here we strengthen the moment condition to $E|X_{j1}|^4 < \infty$ for $j = 1, \ldots, k$, and as before we assume that X_{j1} is distributed on a lattice $x_j + \nu\, e_j$, for integers ν, where e_j is the maximal lattice edge width.

Refinement of theorem about effect of discontinuities – (2)

Theorem. If, for some pair j_1, j_2 with $1 \le j_1 < j_2 \le k$, the ratio $\rho_{j_1 j_2} = (e_{j_2} n_{j_1})/(e_{j_1} n_{j_2})$ satisfies

$$|\rho_{j_1 j_2} - \rho_0| = O\big(n^{-(1/2)\,\{1+(1/\eta)+\delta\}}\big) \tag{4}$$

for some $\delta > 0$, where ρ_0 is an irrational number of type η, then, for each $\delta > 0$,

$$P\left\{\frac{S - E(S)}{(\mathrm{var}S)^{1/2}} \le x\right\} = \Phi(x) + n^{-1/2}\,\tfrac{1}{6}\,\beta\left(1 - x^2\right)\phi(x) + O\big(n^{\delta-(1/2)-(1/2\eta)}\big)\,, \tag{5}$$

uniformly in x.

The theorem is of particular interest when $\eta = 1$, which encompasses almost all irrational numbers (with respect to Lebesgue measure), including all the algebraic irrationals and some transcendental numbers. When $\eta = 1$, (5) is equivalent to

$$P\left\{\frac{S - E(S)}{(\mathrm{var}S)^{1/2}} \le x\right\} = \Phi(x) + n^{-1/2}\,\tfrac{1}{6}\,\beta\left(1 - x^2\right)\phi(x) + O\big(n^{\delta-1}\big)\,, \tag{6}$$

uniformly in x, for each $\delta > 0$.

Result (6) implies that the lattice nature of the distribution of X_{ji} can be ignored, almost up to terms of second order in Edgeworth expansions, when considering the impact of latticeness on the accuracy of normal approximations.

Practical choice of n_1 and n_2

The following assumption is important to the theorem:

$$|\rho_{j_1 j_2} - \rho_0| = O\Big(n^{-(1/2)\,\{1+(1/\eta)+\delta\}}\Big)\,. \tag{4}$$

In practice it is not difficult to choose n_1 and n_2 so that (4) holds. To see how, assume for simplicity that the lattice edge widths e_1 and e_2 are identical, as they would be if (for example) S were equal to a sum or difference of estimators of binomial proportions.

If ρ_0 is an irrational number then the convergents m_1/m_2 of ρ_0 satisfy

$$|(m_1/m_2) - \rho_0| \le m_2^{-2}\,. \tag{7}$$

Therefore, if n_1 and n_2 are relatively prime and n_1/n_2 is a convergent of ρ_0, then (4), for each $\delta \in (0, 3 - (1/\eta)]$, follows from (7).

PRACTICAL CHOICE OF n_1 AND n_2 – (2)

The most difficult case, as far as (4) is concerned, is the one where the convergence rate in (4) is fastest, and arises when $\eta = 1$.

There we need to ensure that

$$|\rho_{j_1 j_2} - \rho_0| = O\big(n^{-1-\delta}\big)$$

for some $\delta > 0$.

Even here there are many options, although the simplest approach is to access the pair (n_1, n_2) from the readily available tables of, or formulae for, convergents for commonly arising irrationals of type 1.

RESULTS FOR THE BOOTSTRAP APPLIED TO MULTI-SAMPLE LATTICE MEANS

Condition (4), stated earlier, and appropriate moment assumptions, are also sufficient for bootstrap versions of the results stated earlier. This includes results about the double bootstrap applied to construct confidence intervals for sums or differences of binomial proportions.

Specifically, provided that

$$|\rho_{j_1 j_2} - \rho_0| = O\Big(n^{-(1/2)\,\{1+(1/\eta)+\delta\}}\Big)\,, \tag{4}$$

where ρ_0 is an irrational number of type 1 (for example, an algebraic irrational),

$$P\left[\frac{S^* - E(S^* \,|\, \mathcal{X})}{\{\mathrm{var}(S^* \,|\, \mathcal{X})\}^{1/2}} \leq x \,\middle|\, \mathcal{X}\right] = \Phi(x) + n^{-1/2}\,\tfrac{1}{6}\,\hat\beta\,\big(1 - x^2\big)\,\phi(x) + n^{-1}\,\Delta(x)\,,$$

$\hat\beta$ denotes the bootstrap estimator of β, and

$$P\Big\{\sup_{-\infty<x<\infty} |\Delta(x)| > n^{\delta}\Big\} = O\big(n^{-C}\big)\,,$$

in which formula $C > 0$ can be taken arbitrarily large, and $\delta > 0$ arbitrarily small, if sufficiently many moments are assumed of the variables X_{ji}.

Numerical properties

Throughout we take $k = 2$ and let X_{ji} be a Bernoulli random variable satisfying $P(X_{ji} = 0) = 1 - P(X_{ji} = 1) = p_j$ for $j = 1, 2$, where $p_1 = 0.4$ and $p_2 = 0.6$. Thus, $\rho_{12} = e_2\, n_1/(e_1\, n_2) = n_1/n_2$, where n_1 and n_2 are the two sample sizes.

We take n_2 to be the integer nearest to $\rho_0\, n_1$, and vary n_1 between 10 and 80; n_1 is plotted on the horizontal axes of each of our graphs. The probability

$$P(x) = P\big[\{S - E(S)\}/(\mathrm{var} S)^{1/2} \le x\big]$$

was approximated by averaging over the results of 10^5 Monte Carlo simulations.

To illustrate the influence of ρ_{12} on the oscillatory behaviour of $P(x)$, and in particular on the accuracy of the normal approximation, each panel in Figures 1 and 2 plots $P(x)$ against n_1 for $x = \Phi^{-1}(\alpha) = z_\alpha$ and $\alpha = 0.95$, 0.85 and 0.75.

Numerical properties – (2)

The first panel of Figure 1 shows results for $\rho_0 = 1$ (indicated by the lines with circles) and $\rho_0 = 2$ (lines with dots), and it is clear that in both cases there is significant oscillatory behaviour.

The second panel of Figure 1 shows that these oscillations decline markedly, and the accuracy of the normal approximation improves considerably, if $\rho_0 = 2^{1/2}$.

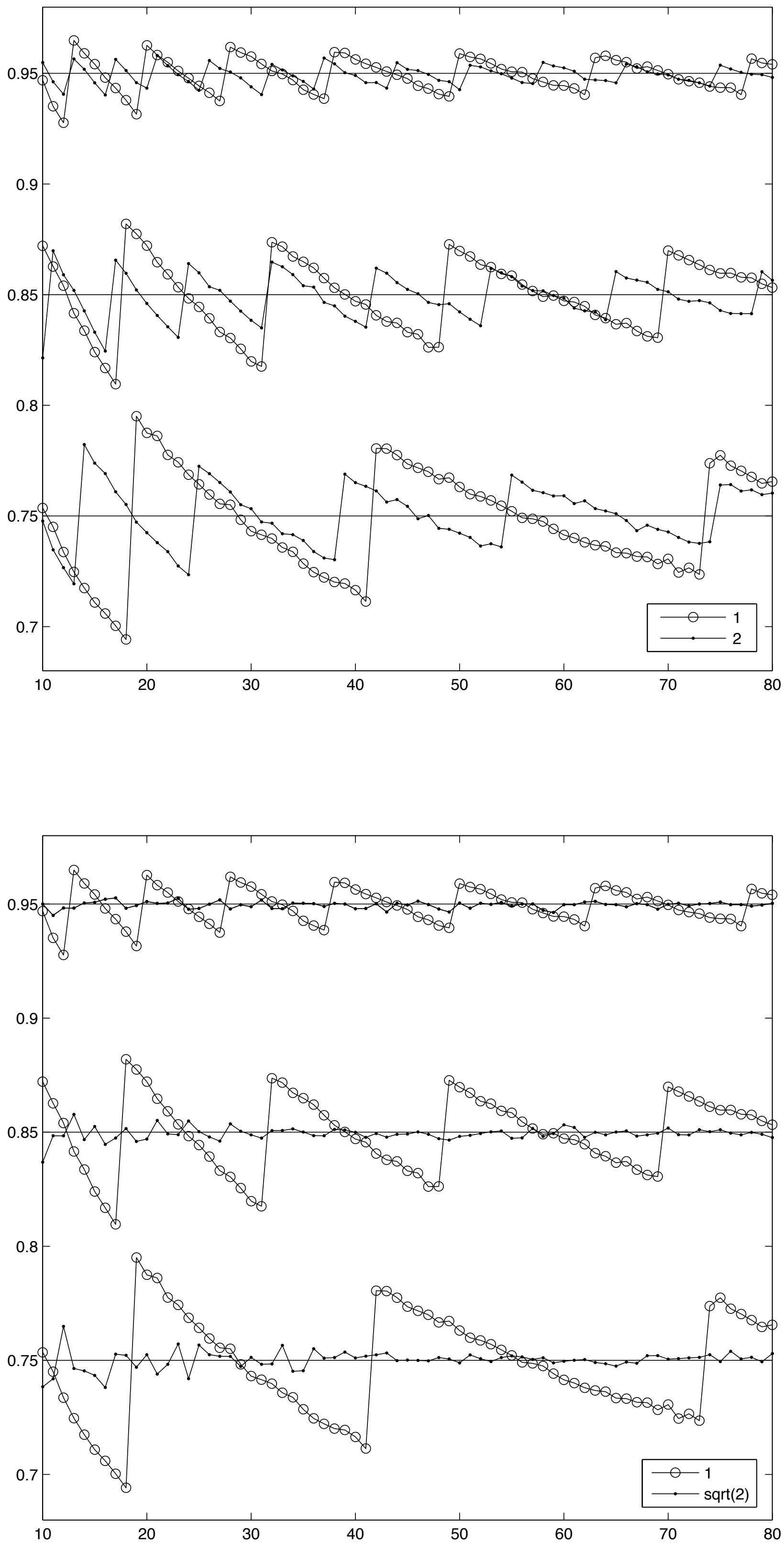


Figure 1: Plots of $P(x)$ against n_1. Plots are given for $x = \Phi^{-1}(\alpha) = z_\alpha$ and $\alpha = 0.95$, 0.85, and 0.75, and for n_2 equal to the nearest integer to $\rho_0 n_1$, with $\rho_0 = 1$ or 2 (in the first panel) and ρ_0= 1 or $2^{1/2}$ (in the second panel).

NUMERICAL PROPERTIES – (3)

Of course, $\rho_0 = 2^{1/2}$ is an algebraic irrational. The first panel of Figure 2 shows that broadly similar values of $P(x)$, although with somewhat more oscillation (reflecting the relatively low upper bounds given in Theorem 1), are obtained for $\rho_0 = \pi/2$, a transcendental irrational whose type is bounded above by 6.61.

The second panel of Figure 2 addresses the theoretical result that there can be less oscillatory behaviour when ρ_{12} converges slowly to a rational number than when it converges quickly.

We consider the cases $n_2 = n_1 + [n_1^{1/5}]$ and $n_2 = n_1 + [n_1^{3/5}]$, where $[x]$ denotes the integer nearest to x. In the first case, ρ_{12} converges relatively quickly to 1, and in the second case the convergence is relatively slow.

Figure 2 demonstrates that, as anticipated, the oscillatory behaviour is less pronounced, and the normal approximation better, in the "slow" case.

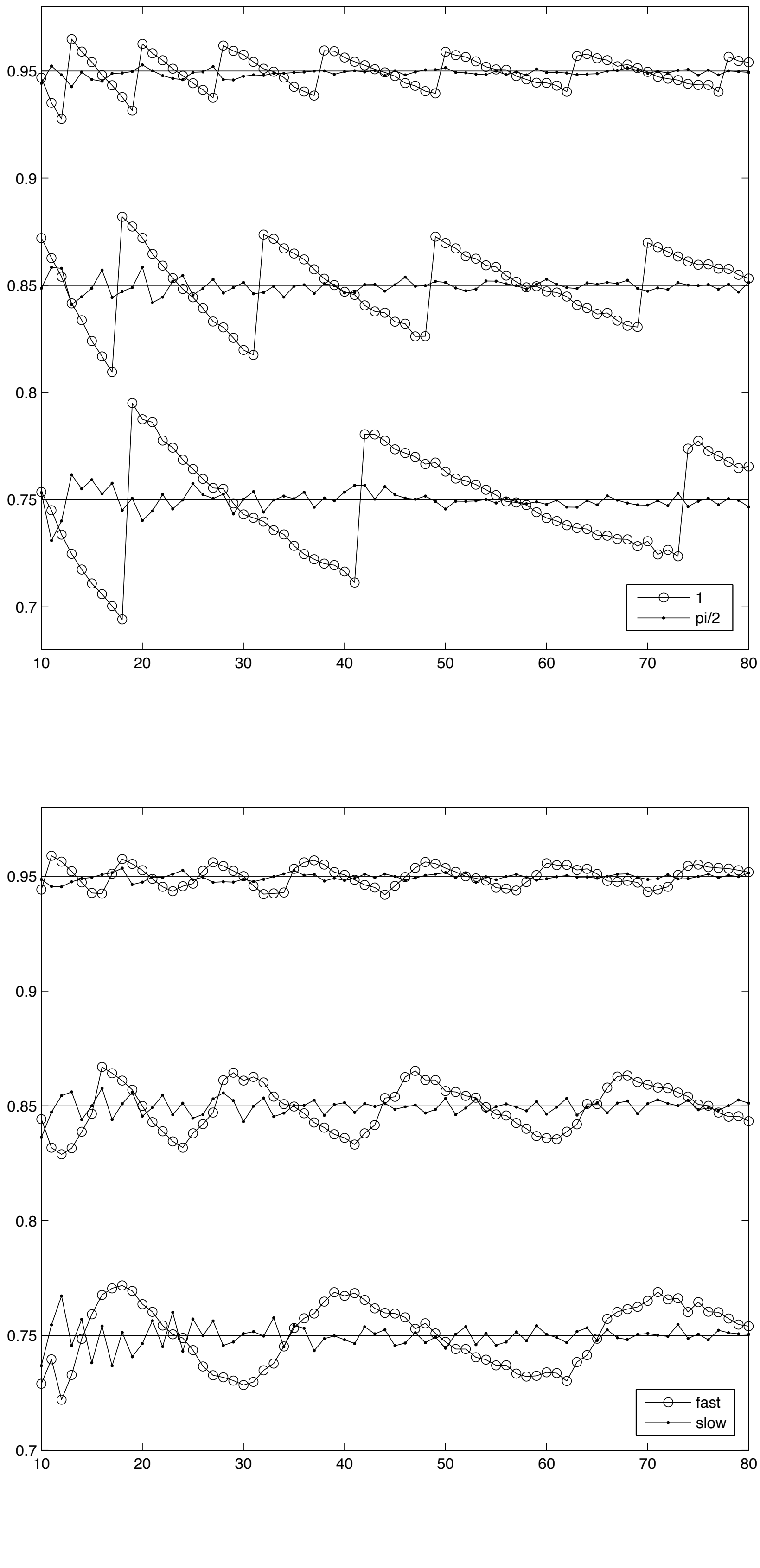


Figure 2: Plots of $P(x)$ against n_1. Plots are as for Figure 1, except that ρ_0= 1 or $\pi/2$ (in the first panel), and ρ_0 converges to 1 rapidly or slowly (in the second panel; see text for details).

Numerical properties — (4)

Figure 3 shows that broadly similar results are obtained in the bootstrap setting. In the figure we give plots of percentile bootstrap estimators of $P(x)$ against n_1, for $x = \Phi^{-1}(\alpha)$ and $\alpha = 0.95$.

Each panel depicts the case $\rho_0 = 1$, and successive panels also give results when $\rho_0 = 3^{1/2}$, $5^{1/2}$, e and $\phi = (1 + 5^{1/2})/2$, respectively.

Each of these values of ρ_0 is an irrational of type 1, and in each instance the oscillations are markedly less, and the normal approximation markedly improved, relative to the case $\rho_0 = 1$.

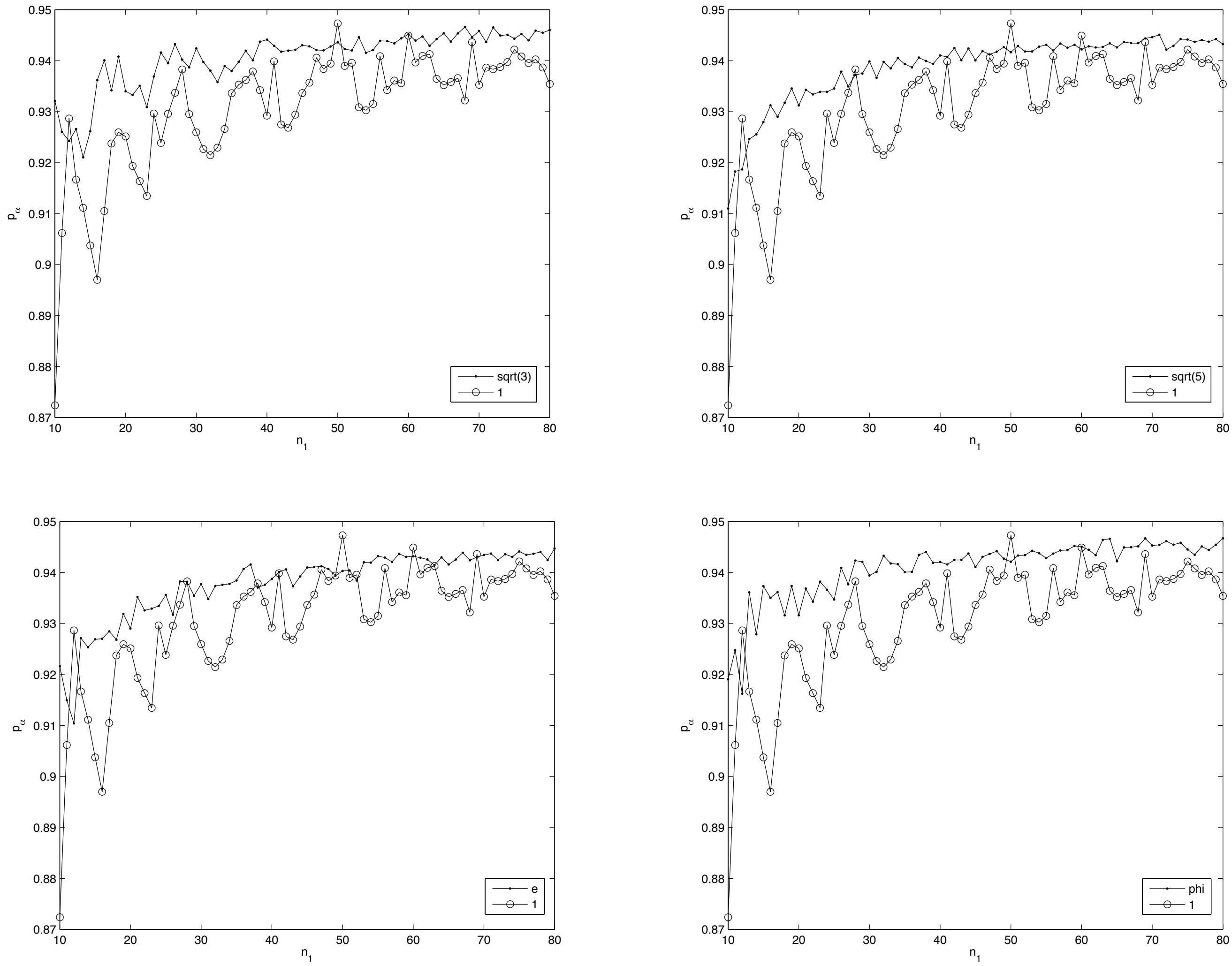


Figure 3: Plots of percentile bootstrap estimators of $P(x)$, against n_1. Plots in each panel show the case ρ_0= 1 and also, in respective panels, the cases $\rho_0 = 3^{1/2}, \rho_0 = 5^{1/2}, \rho_0 = e$ and $\rho_0 = (1 + 5^{1/2})/2$. Throughout, $x = \Phi^{-1}(\alpha)$ where $\alpha = 0.95$.