CONFIDENCE INTERVALS FOR P FROM NORMAL SAMPLES

D. M. Lansky, G. Casella, C. E. McCulloch

BU-998-M Cornell University December, 1988

ABSTRACT

Social scientists report that their colleagues and students frequently misinterpret the meaning of the p-value, particularly when comparing experiments with different sample sizes. This confusion motivates development of a confidence interval for the estimand, p, of the data-based p-value. An interval estimate will have an interpretation that does not change with sample size. One-sided intervals, which include zero, are suggested to address "evidence against H_0." For z and t tests, the distribution of the p-value is derived for all sample sizes. Using these distributions, confidence intervals for p, along with Taylor series approximations, are then constructed.

1. INTRODUCTION

P-values are often considered to be "a way of providing a quantification of strength of evidence" (Kempthorne and Folks 1971, p 314). One widely accepted interpretation of a significance test (where $\hat{p}$ is the observed p-value) is:

p-value	Interpretation
$\hat{p} \leq 0.01$	Strong evidence against the null hypothesis
$0.01 < \hat{p} \leq 0.05$	Moderate evidence against the null hypothesis
$0.10 < \hat{p}$	Little or no real evidence against the null hypothesis.

Several statisticians have suggested that the p-value is not a good measure of evidence against the null hypothesis. Most point out that the meaning of the p-value changes as the sample size changes. Others note that p-values disagree with measures that compare the posterior probability (Bayesian) or likelihood of H_0 to that of H_1 in broad classes of situations and therefore the p-value must be a poor measure of evidence against H_0 (Berger and Sellke, 1987; Johnstone, 1986; Lindley, 1957). This latter criticism focuses on too narrow a question; p-values measure evidence against H_0 in a different sense than do Bayesian or likelihood ratio methods. To interpret a p-value as evidence against H_0 the alternative hypothesis need only be vaguely specified.

The changes in the meaning of the p-value with increasing sample size causes serious problems. For this discussion, higher confidence in a p-value means that one is more certain that there is a meaningful difference between populations and not that the p-value has a lower variance. Logically, there are three possible changes in the interpretation of, or confidence in, p-values with changing sample size: higher confidence in p-values from small samples, higher confidence in p-values from large samples, and equal confidence in p-values from large and small

samples. Royall (1986) and others, including Minturn et al. (1972), correctly point out that, for a given p-value, one ought to have more confidence that the difference between the populations is important with a small sample size than with a large sample size. The fact that many practicing psychological researchers choose an incorrect interpretation (Minturn et al., 1972 and Nelson et al., 1986) only emphasizes Royall's conclusion:

> We should not be surprised to find that despite their apparent inconsistencies, interpretations can be given of all three statements ... that make each of them correct. ... No wonder many good students and scientists find the statistical concepts embodied in "simple" tests of significance elusive.

The evidence against H_0 provided by Bayesian posterior probabilities is different from that provided by p-values for point (or small interval) null hypotheses (Berger and Sellke, 1987). Yet, for some classes of prior densities with one-sided hypotheses, Bayesian and p-value measures of evidence against the null hypothesis agree (Casella and Berger, 1987). These arguments also apply to likelihood ratios, since Bayesian posterior distributions are based on likelihood functions. However, as Barnard (1986) points out, p-values can be used when the alternative hypothesis is only vaguely specified and a likelihood function for the alternative hypothesis cannot be written. For example, in a simple one-sample z-test we can use the p-value to test:

H_0: The $X_i \sim N(0,\sigma^2)$, independent and identically distributed (iid), with σ^2 known

H_1: one or more of the conditions in H_0 is not true.

A dataset which generates a small p-value indicates that either a rare event has occured and H_0 is true or that H_0 is not correct. We do not imply that the p-

value is the measure of choice for testing each aspect of this sort of null hypothesis. We stress the fact that the p-value does provide a measure of evidence against H_0 in a sense different from a that provided by a likelihood ratio test.

When we use a likelihood ratio test (or a Bayesian posterior probability) we will typically make some assumptions under H_1 that match those made under H_0, and allow only one or a few changes, for example:

H_0: The $X_i \sim$iid $N(0,\sigma^2)$, with σ^2 known

H_1: The $X_i \sim$iid $N(\mu,\sigma^2)$, with σ^2 known.

If the common assumptions are valid, then the likelihood ratio tests are preferable because they are more specific. Note that, when the common assumptions (equal variance and independence) are valid, z-tests, t-tests and F-tests, which generate p-values, are likelihood ratio tests.

The key then to interpretation of the p-value is consideration of the sample size. Since the p-value is often interpreted without reference to the sample size, there is a strong need for a measure that has a similar meaning to the p-value, which takes account of the effect of the sample size. We propose a method that has these properties for z, t, and F-tests. This approach is based on construction of an interval estimate for the estimand of the p-value. This interval estimate will have a consistent interpretation with changing sample size.

Since the observed p-value, $\hat{p}$, is a monotone transformation of a calculated test statistic, its distribution can be derived from the distribution of the test statistic. Because, $\hat{p}$ is a point estimate, additional information is needed to have

confidence in the reported p-value. Specifically, $\hat{p}$ alone provides no information about n, thus confidence in the reported $\hat{p}$ is questionable. Even when both $\hat{p}$ and n (or $\Sigma(x_i-\bar{x})^2$ in regression) are reported, no practical method is available for formally combining these to make statements about the meaning of $\hat{p}$. Theoretical methods have been proposed (Kiefer, 1977), but these have not seen much practical use.

When we realize that $\hat{p}$ is a point estimate it is reasonable to ask what $\hat{p}$ is estimating. As established by Joiner (1969), the expected value of $\hat{p}$ is $P(T_0 \geq T)$, where T_0 is a random variable (the test statistic) with some distribution under H_0, and T is also a random variable, independent of T_0, with some distribution under H_1. H_1 is the alternative that assumes that θ is the true value of the unknown parameter. The fact that $\hat{p}$ has a Uniform(0,1) distribution under H_0 is well known. To construct confidence intervals for the true p, the distribution of $\hat{p}$ under H_1 is needed.

Lambert and Hall (1982 and 1983) have shown that the asymptotic distribution of $\log(\hat{p})$ is $N(nc,n\tau^2)$, where c and τ^2 are functions of the parameters of sampled distributions. Due to the difficulty in interpreting c and τ^2, their use is suggested for the comparison of tests "as summary measures of test performance" (Lambert and Hall 1982, p 44). The results reported here provide several advantages over their results for z, t and F tests.

We report the distribution of $\hat{p}$, rather than $\log(\hat{p})$, expressed in terms of the parameter p, rather than c and τ^2. The interpretation of the parameter p is far simpler than that of c and τ^2, the Bahadur half-slope and the variance of the

log-transformed p-value, respectively. In addition, we report results for all sample sizes rather than only asymptotic results. For z-tests, the confidence intervals for p, based on the asymptotic distribution of $\hat{p}$, are very conservative. Hence, the exact distribution that we provide is a needed result.

In constructing confidence intervals based on an exact distribution of $\hat{p}$ we need assumptions similar to those needed for likelihood ratio tests. We consider a specific alternative hypothesis and no longer have a fixed α level for the test. The quantity $1\text{-}P(T_0 \geq T)$ is equivalent to the power function averaged over all sizes for the test (Dempster and Schatzoff, 1965). By recognizing that the converse of a confidence interval for p is a confidence interval for the average power, we may gain some insight on how to use this new interval estimate.

There are at least two ways to motivate a derivation of p, both of which lead to the same definition of p. One way is to define p as the expected value of $\hat{p}$, so $\hat{p}$ is an UMVUE when $\hat{p}$ is based on a complete set of statistics. A second way is to ask that p equal the probability that the test statistic under the null hypothesis (a random variable) is larger than the test statistic under the alternative hypothesis (again, a random variable); in notation: $p = P(T_0 > T)$. These two notions are equivalent (Theorem 1). This leads us to define

$$p \equiv P(T_0 > T) \text{ for } T \sim F_\theta \text{ and } T_0 \sim F_0, \text{ with } T \text{ and } T_0 \text{ independent}$$

for a test with a rejection region of the form $\{T | T > k\}$.

We derive a general expression for the probability density function (pdf) for $\hat{p}$ in terms of the cumulative density functions of the test statistic under the null and alterative hypotheses. We also derive and tabulate expressions for p as a

function of the parameters of the sampled population for z, t and F tests. With expressions for p we then establish and tabulate the density function of $\hat{p}$ for each of these tests. Some properties of these density functions that have implications for the interpretation of p-values are then noted and briefly discussed.

The confidence intervals constructed for p are based on pivotal quantities. The upper end point of each interval is reported, along with a set of expressions for a quadratic Taylor series approximation. In addition, other approximation methods are briefly discussed. A set of tables is included for use in constructing approximate confidence intervals.

2. THE DISTRIBUTION OF THE P-VALUE

We now establish the notational and formal setting for Theorem 1. Consider a test of H_0: $\theta \leq 0$ vs. H_1: $\theta > 0$, based on a test statistic T where the critical region has the form $\{T \mid T > k\}$. We use $\hat{p}$ to represent the usual p-value, which is a random variable equal to $1-F_0(T)$. The observed p-value ($\hat{p}_{obs}$) is an observed value of the random varible $\hat{p}=1-F_0(T)$ with $\hat{p}_{obs}= P(T_0 \geq t_{calc})$, and $E(\hat{p})=P(T_0 \geq T)$.

THEOREM 1. Let T have cdf (cumulative density function) $F_\theta(.)$ and T_0 have cdf $F_0(.)$, with T and T_0 independent. For the statistic $\hat{p} = \hat{p}(t) = 1-F_0(t)$, we have

$$E_\theta[\hat{p}] = E_\theta[\hat{p}(T)] = P_\theta(T_0 > T)$$

Proof: Separate T and T_0 by conditioning on T=t and integrating over the support of $F_\theta(.)$. The fact that $1 - F_0(T) = \hat{p}$ gives the result. □

With the equivalence of $P_\theta(T_0>T)$ and $E_\theta[\hat{p}(T)]$ established, it makes sense to define $p \equiv P_\theta(T_0>T)$ when testing a one-sided hypothesis. For a two-sided hypothesis, we would define $p = \min\Big(P_\theta(T_0<T), P_\theta(T_0>T)\Big)$, because we do not know, a priori, on which side of the null distribution the alternative will lie; for any $\theta \neq \theta_0$, p will be smaller than 1/2.

We can derive a general form for the density function of $\hat{p}$ for a test statistic T with a continuous distribution. This pdf, derived in Theorem 2, will be written in terms of the density function of the test statistic T under both H_0 and H_1 as well as the cumulative density function (cdf) and inverse cdf of the test statistic under H_0. Clearly, the distribution of the test statistic involves the actual parameters of the sampled distribution (which we simplify to θ $(=\mu/\sigma)$ and n). With an expression for p as a function of θ and n (Theorem 3) we can reparameterize the density function of $\hat{p}$ as a function of p (Theorem 4).

THEOREM 2. Let $\hat{p} = 1 - F_{k,0}(T)$, and so $T = F^{-1}_{k,0}(1-\hat{p})$, where the test statistic T has cdf $F_{k,\theta}$ with k a vector of degrees of freedom and θ the true value of the unknown parameter. The pdf of $\hat{p}$ is then

$$f(\hat{p};\theta,k) = f_{k,\theta}\Big(F^{-1}_{k,0}(1-\hat{p})\Big) \,/\, f_{k,0}\Big(F^{-1}_{k,0}(1-\hat{p})\Big)$$

where $f_{k,\theta}$ is the derivative of $F_{k,\theta}$.

Proof: The univariate transformation $\hat{p} = 1 - F_{k,0}(T)$, where $\hat{p}$ is considered a random variable, applied to the true distribution of the test statistic, $f_{k,\theta}(T)$, will yield the above result. □

Using Theorem 1, straightforward calculation will yield Theorem 3. For a random variable with an F-distribution, with k and m degrees of freedom and non-centrality parameter λ we will use $F_{k,m,\lambda}(.)$ and $f_{k,m,\lambda}(.)$ to represent the cumulative density function and the density function respectively. Similarly, we use $F^{-1}_{k,m,0}(.)$ to represent the pdf of the inverse cdf of a central F-distribution with k and m degrees of freedom. We also use the same type of notation for the t-distribution, where we will replace F and f by T and t respectively. We will use the same type of notation to denote the inverse cdf of a χ^2 with m degrees of freedom as $\chi^{2\ -1}_{m}(.)$.

Theorem 3. For tests of $H_0: \theta \leq 0$, on normal populations, with $\theta = \mu/\sigma$, $E_\theta(\hat{p}) = p$ is a function of the parameters of the sampled distribution which depends on the type of test being used:

a) for a z-test, $p = E_\theta(\hat{p}) = G(\theta,n) = \Phi\left(-\theta\,(n/2)^{1/2}\right)$ (1)

b) for a t-test with k degrees of freedom,

$$p = E_\theta(\hat{p}) = G(\theta,n,k) = \frac{1}{2} + \frac{\theta\Gamma(k)}{\Gamma(k/2)}\left(\frac{n}{2\pi}\right)^{1/2} \sum_{s=0}^{\infty} \frac{\Gamma(\frac{k+1}{2}+s)\left(\frac{-\theta^2 n}{2}\right)^s}{\Gamma(k+s+1/2)(2s+1)s!}, \qquad (2)$$

c) for an F-test, with k and m degrees of freedom

$$p = E_\theta(\hat{p}) = G(\theta,n,k,m) = \sum_{i=0}^{\infty} \frac{\lambda^i \exp\{-\lambda\}}{i!} E_{Y_0}\left(F_{k+2i,m,0}\left(\frac{k}{k+2i} Y_0\right)\right), \qquad (3)$$

where $Y_0 \sim F_{k,m,0}$ and $\lambda = n\theta^2/2$.

Proof: Apply Theorem 1. See the appendix for the details.

We now use Theorem 2 and Theorem 3 to construct Theorem 4, a list of density functions for $\hat{p}$ from z, t and F tests.

Theorem 4. For tests of $H_0: \theta \leq 0$ vs. $H_1: \theta > 0$, where $\theta = \mu/\sigma$, on normal populations, where $p = G(\theta,n,k)$ and $\theta = H(p,n,k)$, the inverse function of G(.) exists, the pdf of $\hat{p}$, $f(\hat{p};v)$ can be written as a function of p, which will vary with the type of test as follows:

a) for a z-test $f(\hat{p};p) = \exp\left\{\Phi^{-1}(p)\left((2)^{1/2}\Phi^{-1}(\hat{p}) - \Phi^{-1}(p)\right)\right\}$,

b) for a t-test with k degrees of freedom

$$f(\hat{p};p,n,k) = \exp\left\{-\left(H(p,n,k)/2\right)^2\right\}/\Gamma\left((k+1)/2\right)$$

$$\times \sum_{i=0}^{\infty} \Gamma\left((k+i+1)/2\right)\left(H(p,n,k)T_{k,0}^{-1}(\hat{p})\left(2/(k+(T_{k,0}^{-1}(\hat{p}))^2)\right)^{1/2}\right)^i / i!,$$

c) for an F-test with k and m degrees of freedom

$$f(\hat{p};p,n,k,m) = \frac{\Gamma\left(k/2\right)\exp\left\{-n(H(p,n,k))^2/2\right\}}{\Gamma\left((k+m)/2\right)}$$

$$\times \sum_{i=0}^{\infty} \left(nk(H(p,n,k))^2/2\right)^i \Gamma\left(\frac{k+m}{2}+i\right)\left(\frac{F_{k,m}^{-1}(1-\hat{p})}{m+kF_{k,m}^{-1}(1-\hat{p})}\right)^i / \left(\Gamma\left(\frac{k}{2}+i\right)i!\right).$$

Proof: Use a transformation to go from the distribution of the test statistic to the distribution of $\hat{p}$, then use the results of Theorem 3 to reparameterize the distribution in terms of p (see the appendix for the details).

Before examining the effects of sample size, let us establish a few properties of the pdf of $\hat{p}$. For $\hat{p}$ from a z-test, the pdf is monotone and unimodal in $\hat{p}$ (see the appendix for proof). We already know that when $\theta = \theta_0$ the true $p = 1/2$; hence the pdf of $\hat{p}$ is a Uniform(0,1). As θ moves away from zero the density of $\hat{p}$ "piles up" near one end of the range of $\hat{p}$, [0,1] as is illustrated in Figure 1. With these

results in mind, we can examine the changes in the shape of the density functions as we change n, θ or p.

In Figure 2 we see small p-values become more likely as the sample size increases. This illustrates "Lindley's Paradox" (Lindley, 1957). This is, in fact, the situation most researchers face, where the true θ, or shift, is fixed. If the sample size is "too large" one will detect as statistically significant a very small shift, so small that it means little in practice. Another way to view this is "with a large enough sample size, anything is statistically significant." Experienced researchers are familiar with this phenomenon and require observed p-values to be very small before concluding that important or "significant" differences between populations exist in studies with large sample sizes.

Figure 3 compares experiments with different sample sizes and the same p. Here, larger p-values become more likely as the sample size increases. While this appears to be different from the result observed in Figure 2, the difference can be explained by examining what happens when p is fixed. To keep things simple we will use the expression for p as a function of θ for a z-test from Theorem 3, $p=\Phi\left(-\theta\,(n/2)^{1/2}\right)$. When we compare two experiments with the same p (or estimate of p) and different n, we are comparing populations with different θs. To keep the same p, the change in θ must be the same size as the change in $n^{1/2}$. By comparing Figures 1 and 2, we see that a change in θ usually has a much larger effect on the distribution of $\hat{p}$ than does a change in the sample size. Hence, the combined effect of decreasing θ and increasing n so as to keep p fixed makes large $\hat{p}$ values more likely.

Because 1-P(T_0>T) is the average power (averaged over all α), the confidence interval on p (the expected value of $\hat{p}$), could be inverted to give a confidence interval on the average power. This power interpretation may help to clear up the apparent problem with the increasing width of the confidence interval for p when $\hat{p}$ is fixed and n is increased. In that case, we are forcing the shift to decrease at a rate of $n^{1/2}$. With the shift decreasing we would expect the power to decrease.

These considerations all support Royall's conclusion cited earlier in this paper. The fact that many researchers report and interpret experimental results solely in terms of the $\hat{p}$ values is clearly cause for concern.

3. CONSTRUCTION OF CONFIDENCE INTERVALS

For any two-tailed hypothesis test $\hat{p}$ will always be less than or equal to 1/2. For one-tailed hypothesis tests, evidence that the true p value is larger than 1/2 is rarely useful in practice, so we will restrict the discussion to values of p less than 1/2. For these values of p, a reasonable confidence interval is the region between zero and some upper endpoint (for which we will find an expression). For each test we used a pivotal quantity to find an appropriate upper end point for a confidence interval. These upper endpoints are presented in Theorem 5.

Theorem 5. For tests of $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$, where $\theta=\mu/\sigma$, on normal populations, where p=G(θ,n,k) and θ=H(p,n,k), the inverse function of G(.) exists, the upper end of a one-sided nominally 1-α confidence interval for p, based on $\hat{p}$,
a) for a z-test we get $\Phi\left(2^{1/2}\left(\Phi^{-1}(\hat{p}) - \Phi^{-1}(\alpha)\right)\right)$,

b) for a t-test with k degrees of freedom we get

$$G\left(-T_{k,0}^{-1}(\hat{p})\left(\chi_k^{2\,-1}((1-\alpha)^{1/2})/nk\right)^{1/2} - \Phi^{-1}((1-\alpha)^{1/2})/n^{1/2}, n, k\right), \tag{4}$$

with G(.) defined in (2) of Theorem 3,

c) for an F-test with k an m degrees of freedom we get

$$G\left\{\left(\chi_m^{2\,-1}((1-\alpha)^{1/2})\frac{2k}{nm}\left(F_{k,m}^{-1}(1-\hat{p}) - F_{k,m}^{-1}((1-\alpha)^{1/2})\right)\right)^{1/2}, n, k, m\right\}, \tag{5}$$

with G(.) defined in (3) of Theorem 3.

Proof: In each case, we find pivotal quantities to construct a one-sided confidence interval that bounds θ away from zero. Applying the appropriate G(.) from Theorem 3 completes the derivation (see the appendix for the details).

None of the expressions in Theorem 5 are easy to compute in practice. However, this is not an obstacle since we can derive approximations for these functions that will serve quite well. In particular, we can either use standard approximations, or Taylor series approximations to each expression. The first approach is a good fit for the z-test (based on approximations in Abramowitz and Stegun, 1971).

A general expression for a second order Taylor series approximation to a function (of $\hat{p}$), centered at s is

$$f(\hat{p}) \approx f(s) + f'(s)(\hat{p}-s) + f''(s)(\hat{p}-s)^2/2.$$

With a set of tabulated values for f(s), $f'(s)$ and $f''(s)$, one could easily find an approximate upper end point for the confidence interval. We compute $f'(s)$ and $f''(s)$ where f(s) is, in turn, each of the expressions in Theorem 5. We will then choose a set of values of s for each test (along with other parameters such as degrees of freedom and α) and tabulate these constants.

3.1 Taylor approximation to a confidence interval endpoint for z-test

A two-term Taylor series expansion around s for the upper endpoint of a one-sided (including zero) 1-α confidence interval for p, for a $\hat{p}$ from a z-test is:

$$f(\hat{p}) \approx \Phi\Big(g(s)\Big) + \phi(g(s))(\hat{p}-s)/\Big(2^{1/2}\phi\Big[\Phi^{-1}(s)\Big]\Big)$$

$$+ \phi(g(s))\Big(\Phi^{-1}(s) - g(s)\,\phi\Big[\Phi^{-1}(s)\Big]\Big)(\hat{p}-s)^2/\Big(2\sqrt{2}\Big(\phi\Big[\Phi^{-1}(s)\Big]\Big)^2\Big)$$

where $g(s) = \Big(\Phi^{-1}(s) - \Phi^{-1}(\alpha)\Big)/2^{1/2}$.

These constants are listed in Table 1 for various combinations of $\hat{p}$ and α. Note that these are quadratic polynomials which are most accurate at s, so the group with s closest to the observed $\hat{p}$ should be chosen. This approximation is reasonably accurate as illustrated by Figure 4.

3.2 Taylor approximation to a confidence interval endpoint for t-test

For a second order Taylor series approximation to the upper end point of a one-sided, 1-α confidence interval for p, based on $\hat{p}$ from a t-test we start with

$$f(s) = G\Big(-T_{k,0}^{-1}(s)\Big(\chi_k^{2\,-1}((1-\alpha)^{1/2})/nk\Big)^{1/2} - \Phi^{-1}((1-\alpha)^{1/2})/n^{1/2}, n, k\Big),$$

(4) of Theorem 5, with $G(\theta,n,k)$ defined in (2) of Theorem 3.

We then have

$$f'(s) = 2^{-1/2}\Gamma(k)\Big(\chi_k^{2\,-1}((1-\alpha)^{1/2})\Big)^{1/2}\Big(1+(T_{k,0}^{-1}(s))^2/k\Big)^{(k+1)/2}\sum_1 / \Gamma(\tfrac{k+1}{2})$$

$$f''(s) = \Gamma(k)\Gamma(\tfrac{k}{2})\Big(\tfrac{\pi}{2n}\Big)^{1/2}\Big(\chi_k^{2\,-1}((1-\alpha)^{1/2})\Big)^{1/2}\Big(1+\frac{(T_{k,0}^{-1}(s))^2}{k}\Big)^k / \Gamma(\tfrac{k+1}{2})^2$$

$$\times\Big(k^{-1/2}(k+1)T_{k,0}^{-1}(s)\sum_1 - 2n^{-1/2}\Big(\chi_k^{2\,-1}((1-\alpha)^{1/2})\Big)^{1/2}\Big(1+\frac{(T_{k,0}^{-1}(s))^2}{k}\Big)\sum_2\Big)$$

where $\sum_1$ and $\sum_2$ are

$$\Sigma_1 = \sum_{i=0}^{\infty} \frac{\Gamma(i+(k+1)/2)}{\Gamma((k+i+1)/2)} \left(-\frac{n}{2}\right)^i \left(T_{k,0}^{-1}(s) \left(\frac{\chi_k^{2\,-1}((1-\alpha)^{1/2})}{nk} \right)^{1/2} + \frac{\Phi^{-1}((1-\alpha)^{1/2})}{n^{1/2}} \right)^{2i} / i!$$

$$\Sigma_2 = \sum_{i=0}^{\infty} \frac{\Gamma(i+(k+1)/2)}{\Gamma((k+i+1)/2)} \left(-\frac{n}{2}\right)^i \left(-T_{k,0}^{-1}(s) \left(\frac{\chi_k^{2\,-1}((1-\alpha)^{1/2})}{nk} \right)^{1/2} - \frac{\Phi^{-1}((1-\alpha)^{1/2})}{n^{1/2}} \right)^{2i-1} / (i-1)!.$$

TABLE 1

Coefficients for a quadratic Taylor series approximation to the upper end point of a 1-α confidence interval for p, based on $\hat{p}$ from a z-test. The upper end of the confidence interval is approximated by: $a + b(\hat{p}-s) + c(\hat{p}-s)^2$. The first term of the approximation from each set of coefficients is exact at $\hat{p} = s$.

s	α=0.20 a b c	α=0.05 a b c
0.20	0.5000 1.0076 -1.5146	0.2850 2.3278 -17.9010
0.15	0.5548 0.9981 -1.5690	0.3335 2.4934 -19.3467
0.10	0.6221 0.9600 -1.5923	0.3986 2.6464 -20.7631
0.05	0.7150 0.8575 -1.5325	0.5000 2.7352 -21.8110
0.010	0.8531 0.5807 -1.1777	0.6851 2.4353 -20.0068

Using this approximation, we construct Table 2 for use in calculating an approximate upper end point for a 90% confidence interval for p based on $\hat{p}$ from

a one-sided t-test. Some confidence intervals, with their approximations from Table 2, are shown in Figure 5. It is clear that the approximation does a good job, and is therefore reasonable to use in practice. This confidence interval will provide some perspective on the strength of evidence against H_0.

3.3 Taylor approximation to a confidence interval endpoint for F-test.

For a second order Taylor series approximation to the upper end point of a one-sided, 1-α confidence interval for p, based on $\hat{p}$ from an F-test we start with

$f(s) = G\left(\left(\chi^{2\,-1}_m((1-\alpha)^{1/2})\frac{2k}{nm}\left(F^{-1}_{k,m}(1-\hat{p}) - F^{-1}_{k,m}((1-\alpha)^{1/2})\right)\right)^{1/2}, n, k, m\right)$, (5) of Theorem 5, and $G(\theta, n, k, m)$ as defined in (3) of Theorem 3.

We then have

$$f'(s) = \frac{2}{f_{k,m}\left(F^{-1}_{k,m}(1-\hat{p})\right)} \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!}\left(i-\lambda\right) E_{Y_0}\left(F_{k+2i,m}\left(\frac{k}{k+2i}Y_0\right)\right),$$

where $\lambda = \chi^{2\,-1}_m((1-\alpha)^{1/2})\frac{k}{m}\left(F^{-1}_{k,m}(1-\hat{p}) - F^{-1}_{k,m}((1-\alpha)^{1/2})\right)$.

And we also have

$$\begin{aligned} f''(s) = {} & \left(\frac{k\,\chi^{2\,-1}_m((1-\alpha)^{1/2})}{m\,f_{k,m}(\gamma)}\right)^2 \sum_{i=0}^{\infty} (i^2 - i - \lambda i + 1) E_{Y_0}\left(F_{k+2i,m}\left(\frac{k}{k+2i}Y_0\right)\right) \lambda^{i-2} \frac{\exp\{-\lambda\}}{i!} \\ & + \frac{1}{2}\left\{ \frac{\delta - 1}{\left(\delta f_{k,m}(\gamma)\right)^2} - \frac{\left((k-2)/(1+k\gamma/m) - k(k+m)/\gamma\right)\Gamma\left((k+m)/2\right)(k/m)^{m/2}\,\gamma^{(k-4)/2}}{\left(f_{k,m}(\gamma)\right)^3 \delta\,\Gamma(k/2)\Gamma(m/2)(1+k\gamma/m)^{(k+m-2)/2}} \right\} \\ & \times \sum_{i=0}^{\infty} (i-\lambda) E_{Y_0}\left(F_{k+2i,m}\left(\frac{k}{k+2i}Y_0\right)\right) \lambda^i \exp\{-\lambda\}/i! \end{aligned}$$

check this!

with $\gamma = F^{-1}_{k,m}(1-\hat{p})$ and $\delta = F^{-1}_{k,m}(1-\hat{p}) - F^{-1}_{k,m}((1-\alpha)^{1/2})$.

TABLE 2.
Coefficients for a quadratic Taylor series approximation to the upper end point of a 90% confidence interval for p, based on $\hat{p}$ from a one-sample t-test. The approximate endpoint is $a+b(\hat{p}-p')+c(\hat{p}-p')^2$, where p' is the observed p-value, $\hat{p}$ is from the table and a, b and c are the three values listed in the table in descending order. Note that a is the upper end of the confidence interval for $\hat{p}$. Where values do not appear there were underflow problems in the calculations.

df\p̂	0.001	0.005	0.01	0.025	0.05	0.10	0.15
2				0.0150	0.0432	0.1571	0.3096
				0.8280	1.4841	2.9086	3.0061
				10.1848	16.1911	7.5175	-3.5490
5	0.0008	0.0077	0.0213	0.0775	0.1798	0.3509	0.4774
	1.0909	2.2343	3.1184	4.0987	3.9362	2.9252	2.1842
	201.6135	111.7152	68.2606	9.5668	-9.4899	-8.9802	-6.0169
10	0.0060	0.0384	0.0769	0.1712	0.2855	0.4389	0.5436
	7.4113	8.0806	7.2882	5.4581	3.8794	2.4648	1.7865
	624.2396	-65.5193	-79.6328	-45.2819	-22.1761	-9.1181	-5.0596
15	0.0164	0.0683	0.1166	0.2186	0.3305	0.4735	0.5693
	15.7584	11.0375	8.5707	5.5530	3.6883	2.2662	1.6344
	-1002.7507	-353.0945	-175.9914	-59.7918	-23.7530	-8.6676	-4.6276
20	0.0271	0.0903	0.1429	0.2467	0.3561	0.4928	0.5838
	21.8488	12.4389	9.0524	5.5133	3.5570	2.1541	1.5505
	-3036.0215	-528.1685	-223.1527	-65.1843	-24.0188	-8.3544	-4.3852
30	0.0443	0.1192	0.1751	0.2795	0.3851	0.5145	0.6001
	28.9986	13.6793	9.3907	5.4067	3.3955	2.0281	1.4571
	-6044.8925	-709.5909	-266.3832	-69.1661	-23.9247	-7.9712	-4.1161
40	0.0566	0.1372	0.1945	0.2985	0.4017	0.5269	0.6094
	32.8806	14.2178	9.4970	5.3232	3.2991	1.9568	1.4043
	-7909.7293	-799.8845	-286.0504	-70.5986	-23.7298	-7.7447	-3.9657
50	0.0657	0.1497	0.2076	0.3111	0.4127	0.5351	0.6156
	35.3012	14.5113	9.5378	5.2608	3.2341	1.9096	1.3694
	-9146.8524	-853.7634	-297.2110	-71.2706	-23.5574	-7.5931	-3.8675
75	0.0807	0.1692	0.2279	0.3304	0.4293	0.5475	0.6250
	38.6679	14.8670	9.5611	5.1583	3.1348	1.8389	1.3170
	-10957.1317	-925.9994	-311.4762	-71.9349	-23.2468	-7.3636	-3.7216
100	0.0900	0.1808	0.2398	0.3416	0.4390	0.5547	0.6304
	40.4614	15.0314	9.5571	5.0954	3.0770	1.7982	1.2868
	-11964.5529	-963.2259	-318.4978	-72.1612	-23.0459	-7.2312	-3.6386

4. ASYMPTOTIC INTERVALS

The asymptotic distribution of $\hat{p}$ from a z-test derived by Lambert and Hall (1982) is described by $\ln(\hat{p}) \sim AN(-n\theta^2/2, n\theta^2)$, where AN denotes asymptotically normal. We can derive an approximate confidence interval for p from this asymptotic distribution. Standardizing $\ln(\hat{p})$ we get,

$n^{1/2}\left(\ln(\hat{p})/n\theta + \theta/2\right) \sim AN(0,1)$, and hence

$$1-\alpha = \lim_{n\to\infty}\left(P\left(n^{1/2}\left(\frac{\ln(\hat{p})}{n\theta}+\frac{\theta}{2}\right)<\Phi^{-1}(1-\alpha)\right)\right) = \lim_{n\to\infty}\left(P\left(\frac{\ln(\hat{p})}{n\theta}+\frac{\theta}{2}-\frac{\Phi^{-1}(1-\alpha)}{n^{1/2}}<0\right)\right).$$

When we multiply each term by θ, we must address the fact that θ can be positive or negative. Thus we get

$$1-\alpha = \begin{cases} \lim_{n\to\infty}\left(P\left(\frac{n\theta^2}{2} - \theta n^{1/2}\Phi^{-1}(1-\alpha) + \ln(\hat{p}) < 0\right)\right) & \text{for } \theta<0 \\ \lim_{n\to\infty}\left(P\left(\frac{n\theta^2}{2} - \theta n^{1/2}\Phi^{-1}(1-\alpha) + \ln(\hat{p}) > 0\right)\right) & \text{for } \theta>0 \end{cases}.$$

Solving the quadratic equation for $\theta n^{1/2}$, we get the interval

$$\Phi^{-1}(1-\alpha) - \left(\left(\Phi^{-1}(1-\alpha)\right)^2 - 2\ln(\hat{p})\right)^{1/2} < \theta n^{1/2} < \Phi^{-1}(1-\alpha) + \left(\left(\Phi^{-1}(1-\alpha)\right)^2 - 2\ln(\hat{p})\right)^{1/2}.$$

Note that the left end of this interval is always non-positive, and the right end is always non-negative. Hence, the interval does not exclude any reasonable values of θ. By dividing each term by $2^{1/2}$ then taking each term as an argument of $\Phi(\cdot)$, we get an asymptotic $1-\alpha$ confidence interval for the true p.

Using the fact that $\hat{p} = \Phi(-\overline{X}n^{1/2})$, the approximate coverage probability is

$$P\left(\Phi(-\overline{X}n^{1/2}) < \exp(n^{1/2}\theta\Phi^{-1}(1-\alpha) - n\theta^2/2)\right).$$

Because $\overline{X}n^{1/2} \sim N(\theta n^{1/2}, 1)$, we express the coverage probability for the confidence interval for p based on the asymptotic distribution of $\hat{p}$ as

$$\Phi\left(\theta n^{1/2} + \Phi^{-1}\left(\exp\{n^{1/2}\Phi^{-1}(1-\alpha) - n\theta^2/2\}\right)\right).$$

This interval is very conservative, as is shown by the calculated coverage

probabilities in Table 3. Hence, the exact confidence interval that we have developed gives a useful improvement.

TABLE 3
Coverage probabilities for 95% confidence intervals based on the asymptotic distribution of $\hat{p}$ at various combinations of the shift (θ) and sample size (n). Note that these intervals are all very conservative. Where values do not appear there were underflow problems in the calculations.

θ\n	10	50	100	500	1000	5000	10000
0.1	1.000	1.000	1.000	1.000	1.000	0.992	0.985
0.2	1.000	1.000	1.000	0.999	0.994	0.979	0.973
0.3	1.000	1.000	1.000	0.993	0.986	0.972	1.000
0.4	1.000	1.000	1.000	0.988	0.981	0.985	
0.5	1.000	1.000	0.997	0.983	0.977	1.000	
0.6	1.000	0.999	0.995	0.980	0.974		
0.7	1.000	0.998	0.992	0.977	0.972		
0.8	1.000	0.996	0.990	0.975	0.970		
0.9	1.000	0.994	0.987	0.973	0.990		
1.0	1.000	0.992	0.985	0.971	1.000		
1.1	1.000	0.990	0.984	0.970	1.000		
1.2	1.000	0.989	0.982	0.969	1.000		
1.3	0.999	0.987	0.981	0.998			
1.4	0.999	0.986	0.979	1.000			
1.5	0.998	0.984	0.978	1.000			

5. SUMMARY

P-values are shown to be unbiased estimators of $P(T_0 > T)$, where T_0 is a random variable (the test statistic) under the null hypothesis, and T is the actual test statistic. The p-value is thus a useful measure of evidence against a specified null hypothesis. This use of the p-value does not depend on the specification of an alternative hypothesis. Hence, p-values measure something qualitatively different from what is measured by a likelihood ratio, or a Bayesian posterior probability.

This fact partially explains why many people have found that p-values and Bayesian posterior probabilities often provide conflicting evidence against H_0.

As sample size increases, the meaning of the evidence against a specific null hypothesis provided by a given p-value changes. This fact alone is sufficient motivation to develop a measure, similar to the p-value in meaning, that has an interpretation that does not change with the sample size. We argue that a confidence interval on $P_\theta(T_0>T)$ will have this desired property. In order to construct this interval we have had to assume a specific alternative hypothesis.

For the cases of z, t and F tests, we have developed confidence intervals for the estimand of the p-value, $P(T_0>T)$. We have constructed one-sided intervals that include zero because a p-value or a confidence interval on p is commonly used to answer the question: "can we reject H_0?" The upper end point of the intervals for each test are presented along with Taylor series approximations.

APPENDIX

A1.1 Derivation of the expressions in Theorem 3.

1. Let $p = E(\hat{p})$, with $\hat{p}$ the p-value from a z-test. Let $X = Z - Z_0$; hence, $X \sim N(\theta,2)$. With $p \equiv P_\theta(Z_0 > Z)$, we get $p = P_\theta(X < 0) = \Phi(-\theta/2^{1/2})$.

2. Let $p = E(\hat{p})$, with $\hat{p}$ the p-value from a t-test with k degrees of freedom. With $p\equiv P_\lambda(Y_0>Y)$, we get $p = P_\theta\left(Z_0 - Z(S_0^2/S_1^2)^{1/2} > 0\right)$ with $Y_0=Z_0/S_0$ and $Y=Z/S_1$ where $Z \sim N(\theta n^{1/2},1)$, $kS_i^2/\sigma^2 \sim \chi_k^2$, and Z_0, S_0, Z and S_1 are all independent. Then $T = S_1^2/S_2^2 \sim F_{k,k}$. Hence,

$$p = \int_0^\infty P_\theta(Z_0 - Zt^{1/2} > 0 | T=t) F_T(t)\,dt$$

$$= \int_0^\infty P_\theta\Big((Zt^{1/2} - Z_0)/(1+t)^{1/2} - \theta(nt/(1+t))^{1/2} < -\theta(nt/(1+t))^{1/2} | T=t\Big) F_T(t)\,dt.$$

Because the random variable on the left side of the inequality is distributed N(0,1) we can write,

$$p = E_T\Big(\Phi\Big(-\theta(n\,T/(1+T))^{1/2}\Big)\Big).$$

Using the fact that if $T \sim F_{k,k}$ then $U = T/(1+T) \sim \text{Beta}(k/2, k/2)$ we can use a transformation and write,

$$p = \int_0^\infty \Phi\Big(-\theta(nt/(1+t))^{1/2}\Big) F_T(t)\,dt = \int_0^1 \Phi\Big(-\theta(nu)^{1/2}\Big) \frac{\Gamma(k)}{\Big(\Gamma(k)\Big)^2} u^{k/2-1} (1-u)^{k/2-1}\,du.$$

Taking the derivative and rearranging gives

$$\frac{dp}{d\theta} = -\Big(\frac{n}{2\pi}\Big)^{1/2} \sum_{s=0}^{\infty} \frac{(-\theta^2 n/2)^s}{s!} \frac{\Gamma(s+(k+1)/2)\,\Gamma(k)}{\Gamma(k/2)\,\Gamma(k+s+1/2)}.$$

Now, integrating both sides, we can solve for p

$$p = -\theta\Big(\frac{n}{2\pi}\Big)^{1/2} \sum_{s=0}^{\infty} \frac{(-\theta^2 n/2)^s}{(2s+1)s!} \frac{\Gamma(s+(k+1)/2)\,\Gamma(k)}{\Gamma(k/2)\,\Gamma(k+s+1/2)} + C.$$

For $\theta = 0$, we know that $p = 1/2$, hence $C = 1/2$. Thus, we have

$$p = \frac{1}{2} - \theta\Big(\frac{n}{2\pi}\Big)^{1/2} \sum_{s=0}^{\infty} \frac{(-\theta^2 n/2)^s}{(2s+1)s!} \frac{\Gamma(s+(k+1)/2)\,\Gamma(k)}{\Gamma(k/2)\,\Gamma(k+s+1/2)}$$

3. Let p be the expected value of $\hat{p}$, with $\hat{p}$ the p-value from a F-test with k and m degrees of freedom. By definition $p = P(Y_0 > Y)$, with Y and Y_0 independent. Conditioning by $Y_0 = y$ and integrating over the distribution of Y_0 to get

$$p = \int_0^\infty P_\lambda(Y < y) f_{k,m}(y)\,dy = E_{Y_0}\Big(F_{k,m,\lambda}(Y_0)\Big).$$

Because $f_{k,m,\lambda}(x) = \sum_{i=0}^{\infty} f_{k+2i,m}\Big(\frac{k}{k+2i}x\Big) P\Big(I=i\Big)$, with $I \sim \text{Poisson}(\lambda)$, we can write

$$p = \sum_{i=0}^{\infty} E_{Y_0}\Big(F_{k+2i,m}\Big(\frac{k}{k+2i}Y_0\Big)\Big) \lambda^i \exp\{-\lambda\}/i!.$$

When we substitute $\lambda = n\theta^2/2$ (Searle, 1971) we have the desired result.

A1.2 Derivation of the expressions in Theorem 4

1) For $\hat{p}$ from a z-test, use the transformation

$\theta = H(\hat{p}) = G^{-1}(\hat{p}) = -\sigma\Phi^{-1}(\hat{p})/n^{1/2}$, where $\theta = \mu/\sigma$. Thus, we have

$$f_{\hat{p}}(\hat{p}|p) = \exp\left\{-\frac{1}{2}\left[\left(\Phi^{-1}(\hat{p})\right)^2 + 2\mu n^{1/2}/\sigma\Phi^{-1}(\hat{p}) + \left(\mu n^{1/2}/\sigma\right)^2\right] + \frac{1}{2}\left[\left(\Phi^{-1}(\hat{p})\right)^2\right]\right\}.$$

Recognizing that $-2^{1/2}\Phi^{-1}(p) = \mu n^{1/2}/\sigma$ leads to

$$f_{\hat{p}}(\hat{p}|p) = \exp\left\{\Phi^{-1}(p)\left[2^{1/2}\Phi^{-1}(\hat{p}) - \Phi^{-1}(p)\right]\right\}.$$

2) For $\hat{p}$ from a t-test, Theorem 2 and the symmetry of the t-distribution give

$$f_{\hat{p}}(\hat{p}|p,k) = \left(t_{k,\theta}(-T_{k,0}^{-1}(\hat{p}))\right)/\left(t_{k,0}(-T_{k,0}^{-1}(\hat{p}))\right).$$

This ratio of density functions (see Searle, 1971) reduces to:

$$f(\hat{p};p,k) = \frac{\exp\{-\theta^2/2\}}{\Gamma\left(\frac{k+1}{2}\right)} \sum_{i=0}^{\infty} \left(\theta(-T_{k,0}^{-1}(\hat{p}))\left(2/(k+(-T_{k,0}^{-1}(\hat{p}))^2)\right)^{1/2}\right)^i \Gamma\left(\frac{k+i+1}{2}\right)/i!.$$

Substituting $H(p,n,k)$ for θ yields the statement in the Theorem.

3) For $\hat{p}$ from an F-test, Theorem 2 gives

$$f(\hat{p};p,k,m) = \frac{f_{k,m,\lambda}\left(F_{k,m}^{-1}(1-\hat{p})\right)}{f_{k,m}\left(F_{k,m}^{-1}(1-\hat{p})\right)}, \text{ where } \lambda = \frac{n\theta^2}{2}.$$

When we take this ratio (these pdfs can be found in Searle, 1971), we get

$$f(\hat{p};p,k,m) = \frac{\Gamma(k/2)\exp\{-\lambda\}}{\Gamma((k+m)/2)} \sum_{i=0}^{\infty} \frac{(\lambda k)^i}{i!} \frac{\Gamma(i+(k+m)/2)}{\Gamma(i+k/2)} \left(\frac{F_{k,m}^{-1}(1-\hat{p})}{m+kF_{k,m}^{-1}(1-\hat{p})}\right)^i.$$

A1.3 The pdf of $\hat{p}$ from a z-test is monotone and unimodal in $\hat{p}$.

From the definition of $f(\hat{p};p)$, we have

$$\frac{df(\hat{p}|p)}{d\hat{p}} = \exp\left\{\Phi^{-1}(p)\left(2^{1/2}\Phi^{-1}(\hat{p}) - \Phi^{-1}(p)\right)\right\}\Phi^{-1}(p)2^{1/2}\phi(\hat{p}).$$

Note that the exponential term is always larger than zero, the $\Phi^{-1}(p)$ term is less than zero for p less than $1/2$ $\left(\text{because } \Phi^{-1}(p) > 0 \text{ for } p > 1/2\right)$, and $\phi(\hat{p})$ is always positive. Thus, for a given p, the slope does not change sign. Hence, the pdf is

monotone. Specifically, for p less than 1/2, the slope of the pdf is always negative. Unimodality follows from monotonicity.

A1.4 Derivation of the expressions in Theorem 5

1) Solving p and $\hat{p}$ for μ and $\overline{X}$ respectively leads to $\overline{X} = -\sigma\Phi^{-1}(\hat{p})/n^{1/2}$ and $\mu = -\sigma\Phi^{-1}(p)(2/n)^{1/2}$. Using the fact that $(\overline{X}-\mu)n^{1/2}/\sigma \sim N(0,1)$, we see that $Z = \Phi^{-1}(p)2^{1/2} - \Phi^{-1}(\hat{p}) \sim N(0,1)$. Thus, Z is a pivotal quantity, by this and the symmetry of the Normal distribution we have:

$$P\left[p < \Phi\left(\left(\Phi^{-1}(\hat{p}) - \Phi^{-1}(\alpha)\right)/2^{1/2}\right)\right] = 1-\alpha.$$

The confidence interval follows directly.

2) The quantities $Q_1 = n^{1/2}(\overline{X}-\mu)/\sigma$ and $Q_2 = kS^2/\sigma^2$ are both pivotal quantities (S^2 is the usual sample variance, with k degrees of freedom), and they are independent. Thus, there exist q_1, q_2, q_3 and q_4 so that,

$$P\left(q_1 < n^{1/2}(\overline{X}-\mu)/\sigma < q_2\right) = 1-\alpha_1 \text{ and } P\left(q_3 < kS^2/\sigma^2 < q_4\right) = 1-\alpha_2,$$

which leads to

$$(1-\alpha_1)(1-\alpha_2) = P\left(\frac{\overline{X}}{\sigma} - \frac{q_2}{n^{1/2}} < \frac{\mu}{\sigma} < \frac{\overline{X}}{\sigma} - \frac{q_1}{n^{1/2}} \; ; \sqrt{\frac{q_3}{k}} < \frac{S}{\sigma} < \sqrt{\frac{q_4}{k}}\right).$$

Substituting $\overline{X}/\sigma = \left(n^{1/2}\overline{X}/S\right)\left(S/(\sigma n^{1/2})\right) = -T_{k,0}^{-1}(\hat{p})\, S/(\sigma n^{1/2})$ yields

$$(1-\alpha_1)(1-\alpha_2) = P\left(\frac{-T_{k,0}^{-1}(\hat{p})S}{\sigma n^{1/2}} - \frac{q_2}{n^{1/2}} < \theta < \frac{-T_{k,0}^{-1}(\hat{p})S}{\sigma n^{1/2}} - \frac{q_1}{n^{1/2}} \; ; \left(\frac{q_3}{k}\right)^{1/2} < \frac{S}{\sigma} < \left(\frac{q_4}{k}\right)^{1/2}\right).$$

To obtain a one sided confidence interval for p, with the confidence region abutting 0, set $q_1 = -\infty$ and therefore $q_2 = \Phi^{-1}(1-\alpha_1)$. Similarly, set $q_3 = 0$ and $q_4 = \chi_k^{2\,-1}(1-\alpha_2)$, (the inverse cdf of a χ^2 with k degrees of freedom). This yields

$$(1-\alpha_1)(1-\alpha_2) = P\left(\theta > \frac{-T_{k,0}^{-1}(\hat{p})S}{\sigma n^{1/2}} - \frac{\Phi^{-1}(1-\alpha_1)}{n^{1/2}} \; ; \frac{S}{\sigma} < \left(\chi_k^{2\,-1}(1-\alpha_2)/k\right)^{1/2}\right)$$

Substituting the upper bound from the right hand inequality into the left hand inequality and letting $1-\alpha=(1-\alpha_1)(1-\alpha_2)$, we obtain:

$$1-\alpha \le P\Big(\theta > -T_{k,0}^{-1}(\hat{p})\Big(\chi_k^{2\,-1}(1-\alpha_2)/(nk)\Big)^{1/2} - \Phi^{-1}(1-\alpha_1)/n^{1/2}\Big).$$

Because $G(\theta,n,k)$ is decreasing in θ, we must reverse the inequality when we apply G to both sides of the inequality. Hence, we have

$$1-\alpha \le P\Big(p < G\Big(-T_{k,0}^{-1}(\hat{p})\Big(\chi_k^{2\,-1}(1-\alpha_2)/nk\Big)^{1/2} - \Phi^{-1}(1-\alpha_1)/n^{1/2},n,k\Big)\Big).$$

If we let $1-\alpha = (1-\alpha_1)(1-\alpha_2)$, and we let $\alpha_1=\alpha_2 = 1-(1-\alpha)^{1/2}$, then substitute into the previous expression we obtain the confidence interval in the statement of the theorem.

3) For $\hat{p}$ from an F-test, $\hat{p}=1-F_{k,m}\left(\frac{SSA/k\sigma^2}{SSE/m\sigma^2}\right)$, where SSA/σ^2 has a non-central $\chi^2_{k,\lambda}$ distribution and SSE/σ^2 has a central χ^2_m distribution, and $\lambda = n\theta^2/2$, with $\theta=\mu/\sigma$. Because we know that $E\big(SSA/\sigma^2\big) = k + n\theta^2/2$, we also know that $\frac{SSA/k\sigma^2 - n\theta^2/2k}{SSE/m\sigma^2} \sim F_{k,m}$; hence, it is a pivotal quantity. We can construct a confidence interval on σ^2 from the distribution of SSE/σ^2, which is also a pivotal quantity. The only difficulty here is that these pivotal quantities are not independent. Because SSE and $\frac{SSA/k}{SSE/m}$ are negatively correlated we can construct the following probability statements

$$(1-\alpha_1)(1-\alpha_2)\le P\Big(|\theta| > \Big(\tfrac{2k}{n}\tfrac{SSE}{m\sigma^2}\Big(F_{k,m}^{-1}(1-\hat{p}) - F_{k,m}^{-1}(1-\alpha_1)\Big)\Big)^{1/2}\,;\,\tfrac{SSE}{\sigma^2} < \chi_m^{2\,-1}(1-\alpha_2)\Big)$$

$$\le P\Big(|\theta| > \Big(\tfrac{2k}{nm}\chi_m^{2\,-1}(1-\alpha_2)\Big(F_{k,m}^{-1}(1-\hat{p}) - F_{k,m}^{-1}(1-\alpha_1)\Big)\Big)^{1/2}\Big).$$

Setting $\alpha_1=\alpha_2=1-(1-\alpha)^{1/2}$, applying $G(.,n,k)$ (from Theorem 3) which is a decreasing function, immediately yields the interval in the theorem.

REFERENCES

Abramowitz, M. and Stegun, I. A., Eds. (1972), *Handbook of Mathematical Functions*, National Bureau of Standards (Applied Mathematics Series 55).

Barnard, G. A. (1986), "Discussion" (of Johnstone, 1986), *The Statistician*, 35:499-502.

Berger, J. O. and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence," *Journal of the American Statistical Association*, 82:112-122.

Casella, G. and Berger, R. L. (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," *Journal of the American Statistical Association*, 82:106-111.

Downton, F. (1973), "The Estimation of P(Y<X) in the Normal Case," *Technometrics*, 15:551-558.

Johnstone, D. L. (1986), "Tests of Significance in Theory and Practice," *The Statistician*, 35:491-504.

Joiner, B. L. (1969), "The Median Significance Level and Other Small Sample Measures of Test Efficacy," *Journal of the American Statistical Association*, 64:971-985.

Kempthorne, O. and Folks, L. (1971), *Probability, Statistics and Data Analysis*, Ames, IA: Iowa State University Press.

Kiefer, J. (1977), "Conditional Confidence Statements and Confidence Estimators," *Journal of the American Statistical Association*, 72:789-808.

Lambert, D. and Hall, W. J. (1982), "Asymptotic Lognormality of *P*-Values," *The Annals of Statistics*, 10:44-64.

Lambert, D. and Hall, W. J. (1983), "Correction to: Asymptotic Lognormality of *P*-Values," *The Annals of Statistics*, 11:348.

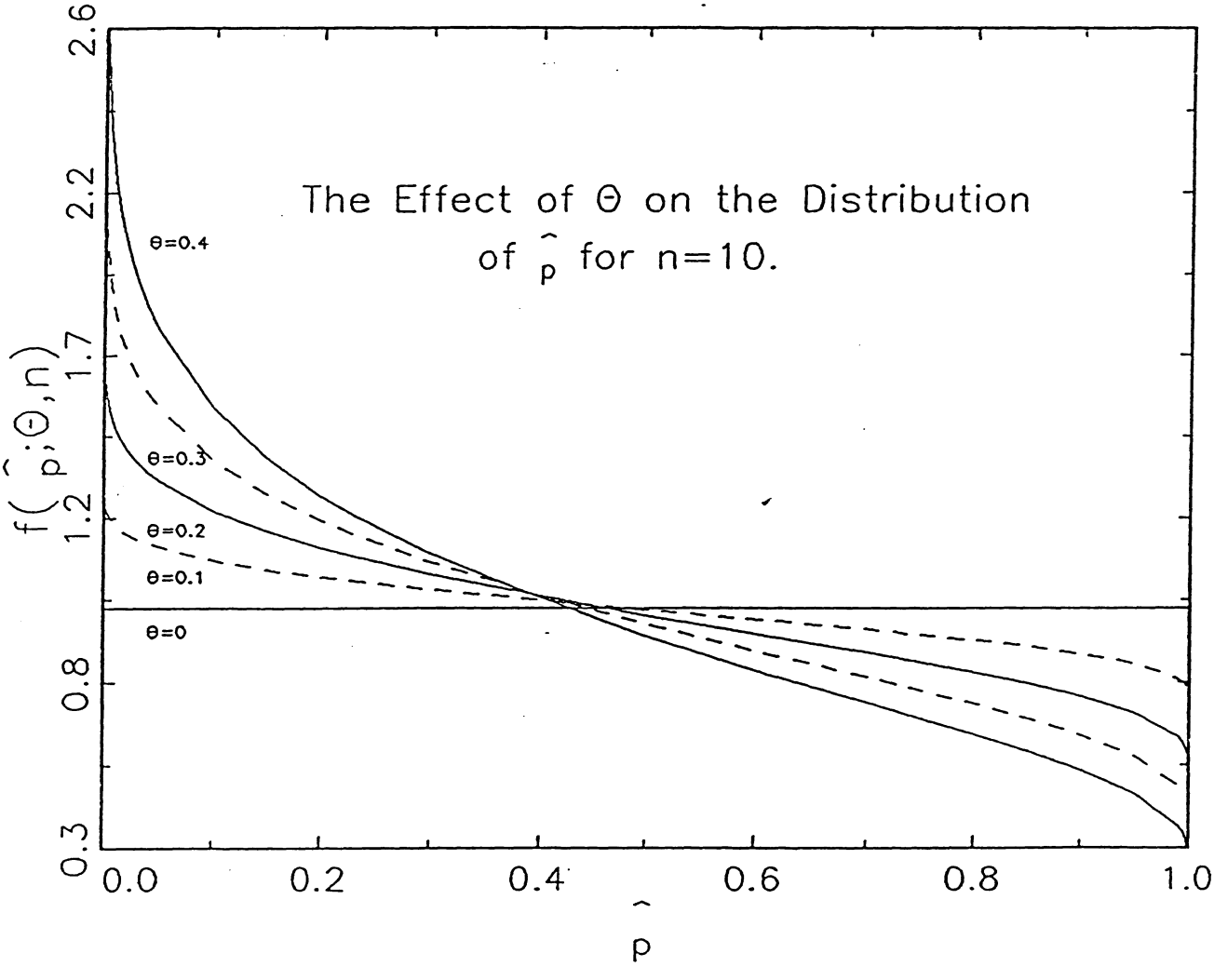
Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44:187-192

Minturn, E. B., Lansky, L. M. and Dember, W. N. (1972), "The Interpretation of Levels of Significance by Psychologists: A Replication and Extension," paper presented at the meeting of the Eastern Psychological Association, Boston.

Nelson, N., Rosenthal, R. and Rosnow, R. L. (1986), "Interpretation of Significance Levels and Effect Sizes by Psychological Researchers," *American Psychologist*, 41:1299-1301.

Royall, R. M. (1986), "The Effect of Sample Size on the Meaning of Significance Tests," *The American Statistician*, 40:313-315.

Searle, S. R. (1971), *Linear Models*, Wiley, N.Y.


The Effect of Θ on the Distribution
of $\hat{p}$ for n=10.
θ=0.4
θ=0.3
θ=0.2
θ=0.1
θ=0
$f(\hat{p};\Theta,n)$
2.6
2.2
1.7
1.2
0.8
0.3
0.0
0.2
0.4
0.6
0.8
1.0
$\hat{p}$

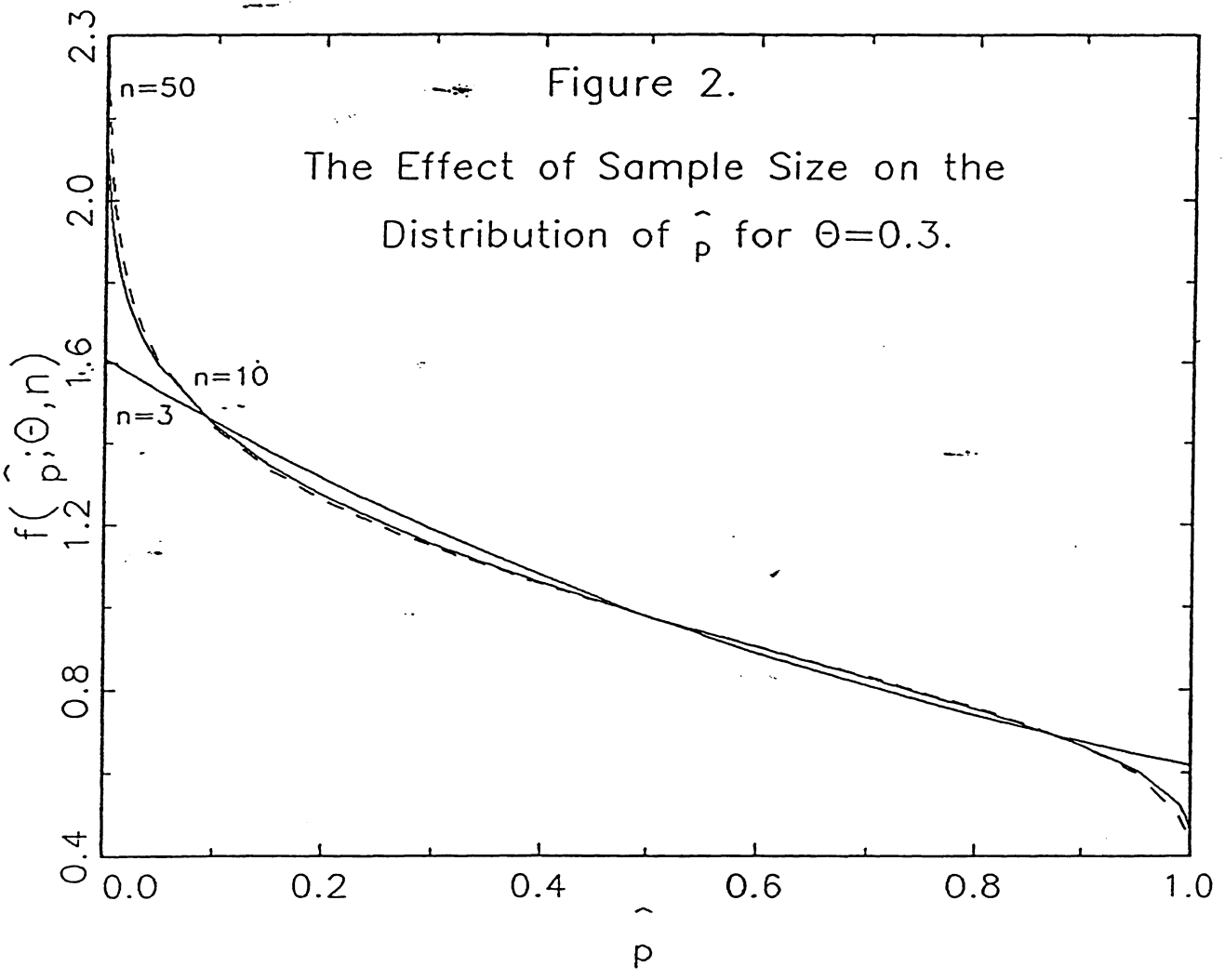


Figure 2.
The Effect of Sample Size on the
Distribution of $\hat{p}$ for $\Theta=0.3$.
n=50
n=10
n=3
$f(\hat{p};\Theta,n)$
$\hat{p}$
0.0
0.2
0.4
0.6
0.8
1.0
0.4
0.8
1.2
1.6
2.0
2.3

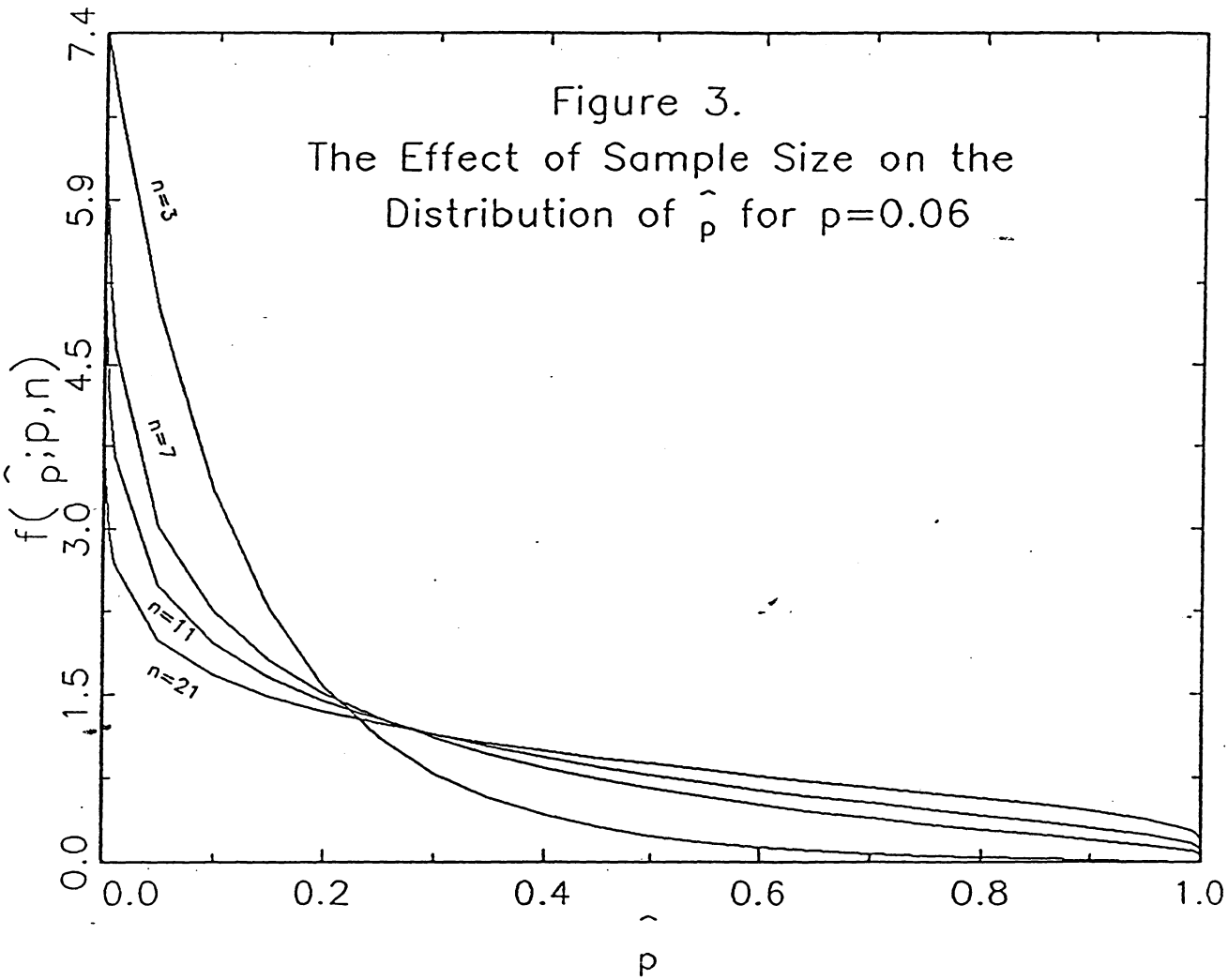


Figure 3.
The Effect of Sample Size on the
Distribution of $\hat{p}$ for p=0.06
n=3
n=7
n=11
n=21
$f(\hat{p};p,n)$
0.0
1.5
3.0
4.5
5.9
7.4
0.0
0.2
0.4
0.6
0.8
1.0
$\hat{p}$

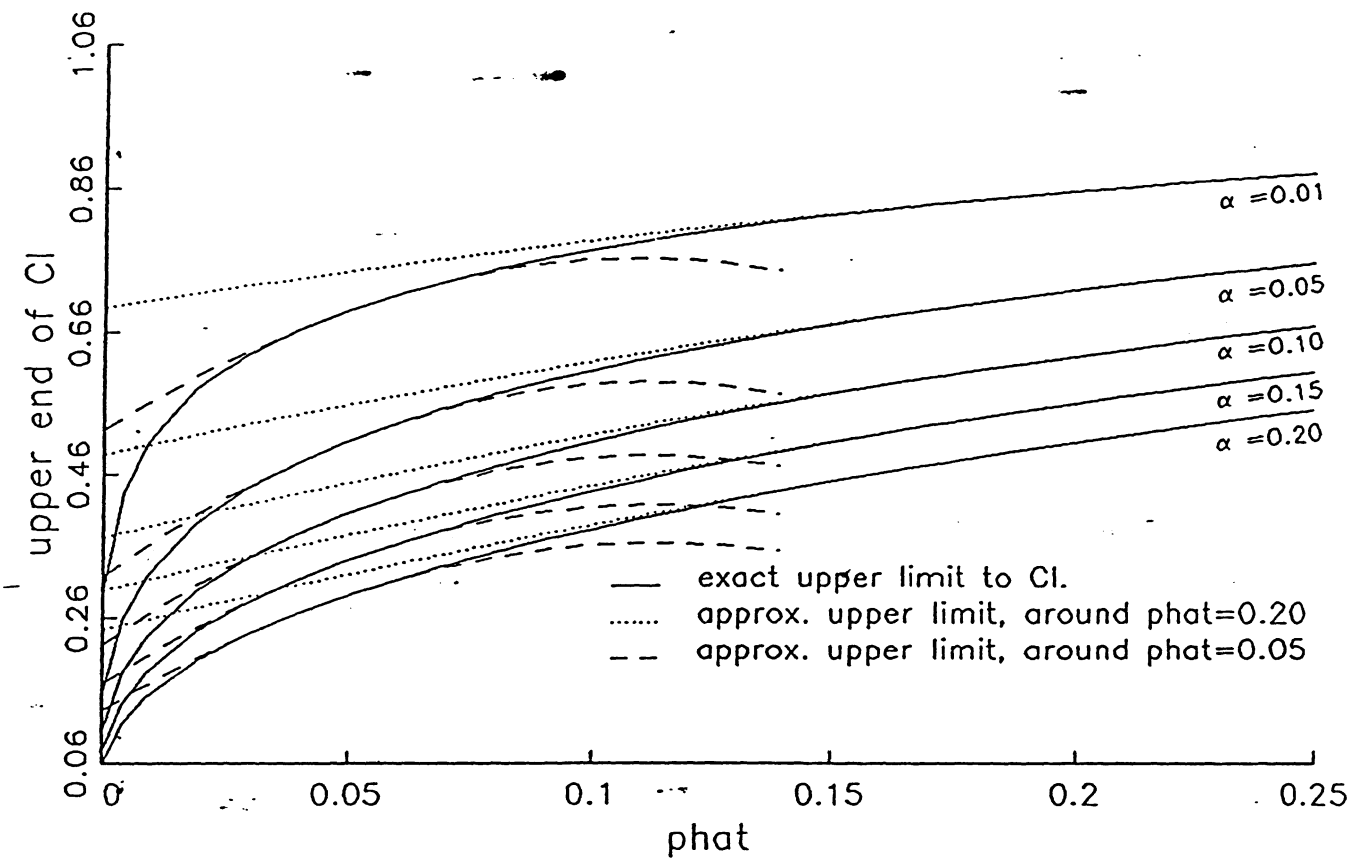


Figure 4. Upper end of Confidence intervals for p based on phat from a z–test, at several levels of alpha.

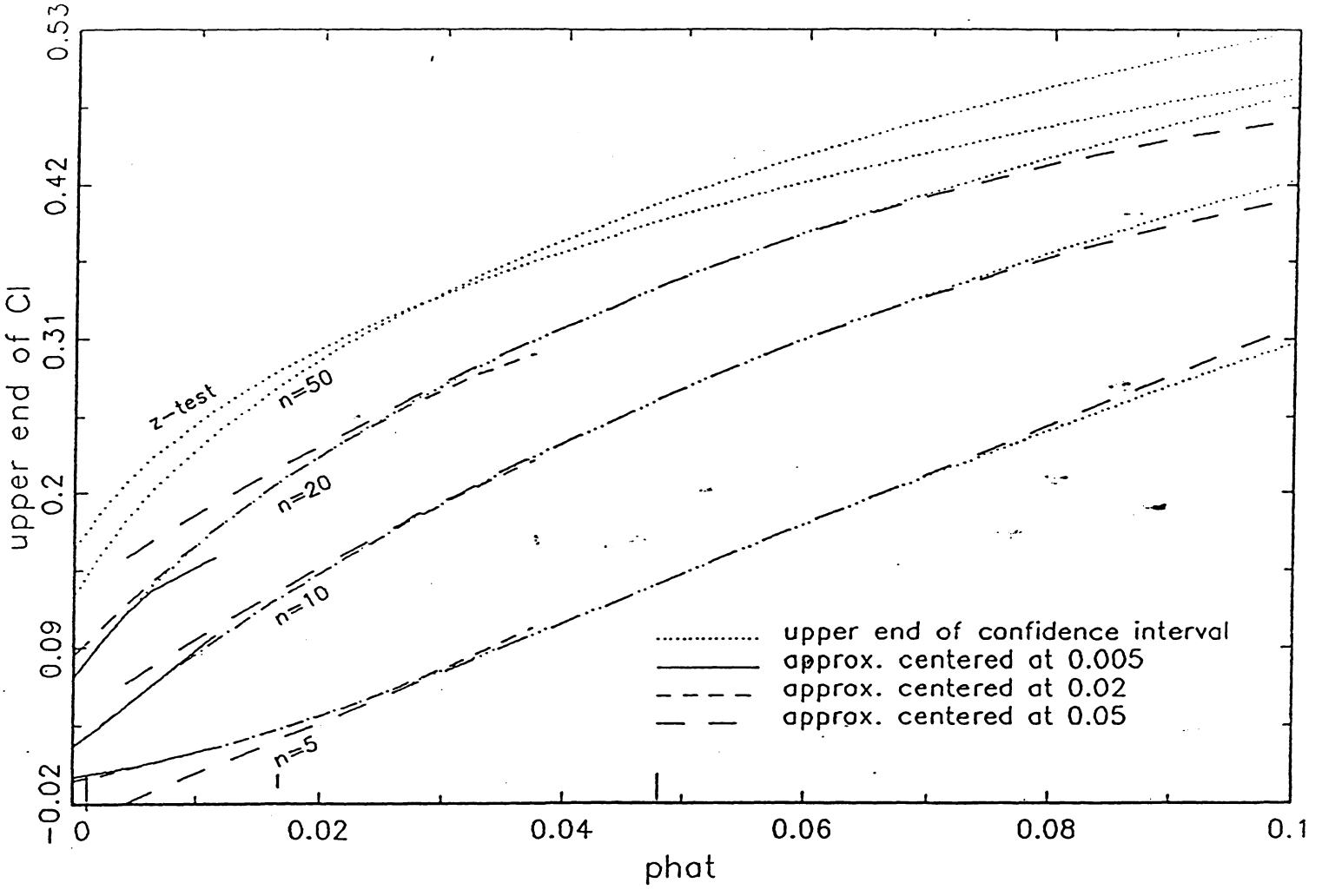


Figure 5: Upper end of 90% Confidence Interval for p, based on phat from a t-test