Ann. Inst. Statist. Math.
Vol. 43, No. 3, 565–578 (1991)

BOUNDS FOR THE SAMPLE SIZE TO JUSTIFY NORMAL APPROXIMATION OF THE CONFIDENCE LEVEL

THOMAS HÖGLUND

Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

(Received December 19, 1988; revised November 10, 1989)

Abstract. The normal approximation of the confidence level of the standard confidence intervals leaves an error of the order $O(1/n)$ (and not only $O(n^{-1/2})$). We use the first order term in the error to obtain simple lower bounds for the sample size.

Key words and phrases: Confidence interval, normal approximation, Edgeworth expansion, bounds for the sample size.

1. Introduction

The interval

$$\bar{x} \pm 1.64s/n^{1/2} \tag{1.1}$$

is a confidence interval for the population mean, μ, with an approximate confidence level of 90%. The approximation may be bad if n is small or if the population has a skewness, and it is therefore desirable to have a lower bound for n. Such bounds exist. Cochran ((1977), p. 42) suggested the bound

$$n > 25g_1^2 \tag{1.2}$$

as a crude measure for the corresponding 95% interval in the case of simple random sampling. Here g_1 is an estimate of the skewness $\gamma_1 = \mu_3/\sigma^3$, σ^2 is the variance and μ_3 stands for the third order central moment. Dalén (1986) proposed a related bound.

Let $F_n(t)$ denote the distribution function of $n^{1/2}(\bar{x} - \mu)/s$, and suppose that F_n admits the Edgeworth expansion

$$F_n(t) = \Phi(t) + n^{-1/2}P(t)\phi(t) + n^{-1}Q(t)\phi(t) + o(n^{-1}). \tag{1.3}$$

Here Φ and ϕ denote the standard normal distribution and density functions, respectively. P and Q are polynomials whose coefficients depend on the distribution

of the sample. In typical cases $P(t) = \gamma_1 H(t)$, where H depends on the sampling procedure, but not on the population. This explains the form of the bound (1.2).

The expansion (1.3) implies

$$F_n(t - n^{-1/2}P(t)) = \Phi(t) + O(n^{-1}). \tag{1.4}$$

If (1.2) is not satisfied one can therefore use the interval

$$\bar{x} - (1.64 - n^{-1/2}\hat{P}(1.64))s/n^{1/2} < \mu < \bar{x} + (1.64 + n^{-1/2}\hat{P}(-1.64))s/n^{1/2} \tag{1.5}$$

instead of (1.1) and get an error in the order of $O(1/n)$. Here $\hat{P}$ is an estimate of P. This has been noted by several authors (see Johnson (1978), Hall (1983) and Abramovitch and Singh (1985)). As explained in the latter two papers mentioned, further corrections can be made.

What is said above must, however, be interpreted with caution:

If we want both tail probabilities to be close to 5%, then a bound of the form Kg_1^2 is the relevant measure, and the correction term in (1.5) will in general improve the approximation compared to (1.1).

If we, however, want the overall confidence level to be close to 90%, then a bound of the form (1.2) is not the relevant measure, and the correction term in (1.5) will not in general improve the approximation compared to (1.1).

The reason as to why the overall confidence level fits better than either one of the two tail probabilities is that in typical situations, the polynomial $P(t)$ is even and therefore the $n^{-1/2}$ terms cancel. This simple observation is, of course, not new. It is made by Hall (1983), for example.

In this paper, we shall consider uncorrected intervals, and use the principal error term as given by the Edgeworth expansion to obtain simple lower bounds for the sample size. These bounds give an error approximately equal to any pre-assigned small number.

Thus for example the bound $n > 31 + 18g_1^2 + g_2$ will, when the interval (1.1) is based on independent observations, give an error which approximately equals 1%. Here g_2 is an estimate of the excess $\gamma_2 = \mu_4/\sigma^2 - 3$, where μ_4 stands for the fourth order central moment. Divide the bound by e if you accept the error e%. The corresponding bound which will give an error of 1/2% in each of the two tail probabilities is $n > 486g_1^2$. Divide the bound by e^2 if you accept the error $e/2$%.

Note that these bounds are not arbitrary but are founded on the theorem in Section 2. We shall also give the corresponding bounds when 1.64 is replaced by the corresponding t-percentile, and thus be able to decide when the t-percentile will give a better approximation than the normal percentile. The bounds can be modified to suit other confidence intervals, other confidence levels, and other sampling procedures as well as discrete distributions as explained in the following sections.

2. Continuous populations

Consider a sample $x = (x_1, \ldots, x_n)$ from a distribution that depends on an unknown parameter θ, and a statistic $T_n(x, \theta)$ whose distribution function F_n admits an expansion of the form (1.3) with P even, and Q odd; $P(-t) = P(t)$, $Q(-t) = -Q(t)$.

The confidence set

$$I_\alpha = \{\theta;\ -z_{\alpha/2} < T_n(x, \theta) < z_{\alpha/2}\}, \tag{2.1}$$

where $\Phi(z_\alpha) = 1 - \alpha$, has the confidence level

$$1 - \alpha_n = F_n(z_{\alpha/2}) - F_n(-z_{\alpha/2}).$$

Introduce the two tail probabilities

$$\alpha_n^+ = 1 - F_n(z_{\alpha/2}), \qquad \alpha_n^- = F_n(-z_{\alpha/2}).$$

Then $\alpha_n = \alpha_n^+ + \alpha_n^-$. If we use (1.3) and the fact that P is even, we conclude

$$\alpha_n = \alpha + \delta_\alpha/n + o(1/n), \qquad \alpha_n^+ = \alpha/2 + \beta_\alpha/n^{1/2} + O(1/n),$$

where

$$\delta_\alpha = -2Q(z_{\alpha/2})\phi(z_{\alpha/2}), \qquad \beta_\alpha = -P(z_{\alpha/2})\phi(z_{\alpha/2}).$$

Therefore δ_α/n is a measure of the error in the approximation of α_n. Neglecting terms of smaller order than $1/n$ we conclude that the bound

$$n > |\delta_\alpha|/\epsilon \tag{2.2}$$

will produce an error less than ϵ. The corresponding bound for each of the two tail probabilities is

$$n > \beta_\alpha^2/\epsilon^2. \tag{2.3}$$

THEOREM 2.1. *Assume that F_n admits the Edgeworth expansion* (1.3) *with P even and Q odd. Let $n(\alpha, \epsilon) = [|\delta_\alpha|/\epsilon]$ and $n^+(\alpha, \epsilon) = [\beta_\alpha^2/\epsilon^2]$. If $\delta_\alpha \neq 0$, then $|\alpha_{n(\alpha,\epsilon)} - \alpha| = \epsilon + o(\epsilon)$ as $\epsilon \to 0$. If $\beta_\alpha \neq 0$, then $|\alpha^+_{n^+(\alpha,\epsilon)} - \alpha/2| = \epsilon + O(\epsilon^2)$ as $\epsilon \to 0$.*

PROOF. The theorem follows from the above expansions for α_n and α_n^+, and the fact that $n(\alpha, \epsilon) = |\delta_\alpha|/\epsilon + O(1)$ and $n^+(\alpha, \epsilon) = \beta_\alpha^2/\epsilon^2 + O(1)$. □

A weak point here is that if we have the favourable situation of $\beta_\alpha = 0$ then (2.3) takes the optimistic form $n > 0$. In this case the error is measured by the next term in the expansion of α_n^+, that is by $\delta_\alpha/(2n)$. So (2.3) has to be replaced by $n > |\delta_\alpha|/(2\epsilon)$. If $\beta_\alpha \neq 0$, then $\beta_\alpha/n^{1/2}$ is a good approximation of the error only when n is so large that $\beta_\alpha/n^{1/2}$ is big compared to the next term. That

is, when n is large compared to $(\delta_\alpha/2\beta_\alpha)^2$. Similarly, if $\delta_\alpha = 0$ then the error is measured by the next term in the expansion of α_n, which in regular cases has the form ψ_α/n^2. Therefore (2.2) has to be replaced by $n > (|\psi_\alpha|/\epsilon)^{1/2}$. And if $\delta_\alpha \neq 0$, then δ_α/n is a good approximation of the error only when n is large compared to $|\psi_\alpha/\delta_\alpha|$.

The theorem can be used in the following way: choose α and ϵ. Compute, if possible, the sample size $n(\alpha, \epsilon)$. Draw the sample.

A complication here is that $n(\alpha, \epsilon)$ may depend on unknown characteristics of the population such as γ_1 and γ_2. If this is the case, start with a reasonably large sample of size n_0. When choosing n_0 you can use the form of $n(\alpha, \epsilon)$ and what you think about the values of the unknown characteristics. Then, compute $\hat{n}_0(\alpha, \epsilon)$, the bound given by the theorem but with the unknown characteristics replaced by estimators based on the sample. The sample is sufficiently large if $n_0 \geq \hat{n}_0(\alpha, \epsilon)$. Otherwise, enlarge the sample to a sample of size $n_1 \geq \hat{n}_0(\alpha, \epsilon)$. Then compute $\hat{n}_1(\alpha, \epsilon)$ in the same way as $\hat{n}_0(\alpha, \epsilon)$ but where the estimators are based on the enlarged sample. The enlarged sample is sufficiently large if $n_1 \geq \hat{n}_1(\alpha, \epsilon)$. This is probably the case, otherwise we can enlarge the sample a second, a third, ..., a k-th time until $n_k \geq \hat{n}_k(\alpha, \epsilon)$.

This is a background to the following:

General rule. If you accept the error ϵ in the approximation of the confidence level, use the bound $n > |d_\alpha|/\epsilon$. If you only accept the error $\epsilon/2$ in the approximation of each of the tail probabilities, use the bound $n > (2b_\alpha/\epsilon)^2$. Here d_α and b_α are consistent estimators of δ_α and β_α, respectively.

Note that the rule is an attempt to hit the exact sample size. The rule is not designed to be on the safe side.

We shall consider four special cases in some detail. Three of the cases are based on independent observations, but with different statistics T. The remaining case (Case 2) is the analogue of Case 1 when the independent observations are replaced by a simple random sample.

Case 1. $X_1, \ldots, X_n$ are independent and identically distributed, σ is known or a known function of μ, $T = n^{1/2}(\bar{X} - \mu)/\sigma$, and $EX_1^4 < \infty$. Here $\bar{X} = (X_1 + \cdots + X_n)/n$. We shall also need the technical condition $\limsup_{|\xi| \to \infty} |E(\exp i\xi X_1)| < 1$, this is always the case if the distribution of X_1 has an absolutely continuous component.

Case 2. $X_1, \ldots, X_n$ is a simple random sample from a finite population of size N, which is continuous in the sense of Robinson (1978). σ is assumed to be known and $T = n^{1/2}(\bar{X} - \mu)/(\sigma(qN/(N-1))^{1/2})$. Here $q = 1 - p$, where p is the sampling ratio $p = n/N$.

Case 3. $X_1, \ldots, X_n$ are i.i.d., and, $T = n^{1/2}(\bar{X} - \mu)/s$, where $s^2 = \sum(X_i - \bar{X})^2/(n-1)$. The technical condition we shall need is that $EX_1^8 < \infty$, and that the distribution of X_1 has an absolutely continuous component.

Case 4. As Case 1 but with $T = n^{1/2}(\bar{X} - \mu)/\sigma(\bar{X})$ where the standard deviation $\sigma(\mu)$ is a function with the following properties: its domain contains $\bar{X}$ with probability 1 for each n, and it is two times continuously differentiable in the interior of the domain. We shall also assume that μ belongs in the interior of the domain.

Below is a table displaying the bounds for the sample sizes in Cases 1, 3 and 4 for some options of α and ϵ (unit: %). The options for ϵ are examples rather than recomendations. Bounds for other values of ϵ are obtained in the following way: $B(\alpha, \epsilon_2) = \epsilon_1 B(\alpha, \epsilon_1)/\epsilon_2$, $B_+(\alpha, \epsilon_2) = \epsilon_1^2 B_+(\alpha, \epsilon_1)/\epsilon_2^2$. Here $B(\alpha, \epsilon)$ and $B_+(\alpha, \epsilon)$ are the bounds that give the overall error ϵ, respectively the error $\epsilon/2$ in each tail.

A corresponding table valid for discrete distributions can be found in Section 4.

Table 1. Bounds for sample size in Cases 1, 3 and 4.

	α	ϵ	Overall error ϵ	Error $\epsilon/2$ in each tail
Case 1	1	0.5	$\lvert 1.5g_1^2 - 2.3g_2\rvert$	$29g_1^2$
	5	1	$\lvert 2.8g_1^2 - 0.8g_2\rvert$	$31g_1^2$
	10	2	$\lvert 0.7g_1^2 + 0.2g_2\rvert$	$9g_1^2$
Case 3	1	0.5	$\lvert 28 + 45g_1^2 - 4.5g_2\rvert$	$189g_1^2$
	5	1	$\lvert 28 + 25g_1^2 - 1.6g_2\rvert$	$286g_1^2$
	10	2	$16 + 9g_1^2 + 0.4g_2$	$121g_1^2$
Case 4	1	0.5	$\lvert 1.5g_1^2 - 2.3g_2 + 60g_1\sigma'(\bar{x}) - 229\sigma'(\bar{x})^2 + 49\sigma''(\bar{x})\sigma(\bar{x})\rvert$	$(5.4g_1 - 38.4\sigma'(\bar{x}))^2$
	5	1	$\lvert 2.8g_1^2 - 0.8g_2 + 12g_1\sigma'(\bar{x}) - 81\sigma'(\bar{x})^2 + 44\sigma''(\bar{x})\sigma(\bar{x})\rvert$	$(5.5g_1 - 44.9\sigma'(\bar{x}))^2$
	10	2	$\lvert 0.7g_1^2 + 0.2g_2 - 2.3g_1\sigma'(\bar{x}) - 16\sigma'(\bar{x})^2 + 23\sigma''(\bar{x})\sigma(\bar{x})\rvert$	$(2.9g_1 - 27.9\sigma'(\bar{x}))^2$

The corresponding bounds for Case 2 are

$$\min(n, N-n) > \frac{2|c|}{1 + a/N + \sqrt{(1 + a/N)(1 + b/N)}}, \tag{2.4}$$

respectively

$$\min(n, N-n) > \frac{2C}{1 + 4C/N + \sqrt{1 + 4C/N}}. \tag{2.5}$$

Here, C is the corresponding bound given by the last column in Case 1, c is the expression within the absolute value signs in the column for overall error in Case

1, $b = -\text{sign}(c)(2g_2 + 6)d(\alpha, \epsilon)$ and $a = 4|c| + b$. Here $d(1\%, 0.5\%) = 2.26$, $d(5\%, 1\%) = 0.8$ and $d(10\%, 2\%) = -0.21$. We have assumed that N is so large that the expressions under the squareroot sign are positive.

For the statistic in Case 3 the overall error will be smaller for some distributions if we use the t-percentile instead of the normal percentile, $z_{\alpha/2}$.

Table 2. Bounds for sample size in Case 3 when using t-percentile.

α	ϵ	Use t-percentile if	Overall error ϵ	Error $\epsilon/2$ in each tail
1	.5	$39 + 48g_1^2 - 13g_2 > 0$	$\lvert 45g_1^2 - 4.5g_2\rvert$	$189g_1^2$
5	1	$34 + 62g_1^2 - 4g_2 > 0$	$\lvert 25g_1^2 - 1.6g_2\rvert$	$286g_1^2$
10	2	$1 > 0$	$\lvert 9g_1^2 + 0.4g_2\rvert$	$121g_1^2$

The bounds have particularly simple forms when $z_{\alpha/2} = \sqrt{3}$ i.e. when $\alpha = 8\%$. Table 3 gives a general picture and facilitates the comparison between the different bounds.

Table 3. Bounds for sample size when $\alpha = 8\%$.

	Overall error 1%	Error 1/2% in each tail
Case 1	$(5/2)g_1^2$	$35g_1^2$
Case 2	$\dfrac{5g_1^2}{1 + 10g_1^2/N + \sqrt{1 + 10g_1^2/N}}$	$\dfrac{70g_1^2}{1 + 140g_1^2/N + \sqrt{1 + 140g_1^2/N}}$
Case 3	$30 + 20g_1^2$	$430g_1^2$
Case 4	$(5/2)\lvert g_1^2 + 18\sigma(\bar{x})^2(d^2/d\bar{x}^2)\log\sigma(\bar{x})\rvert$	$35(g_1 - 9\sigma'(\bar{x}))^2$

The two-sided bound in Case 3 has to be replaced by $20g_1^2$ when the t-percentile is used.

Both Table 3 and the example in Section 3 reflect the general fact that the bound $n(\alpha, \epsilon)$ is of a smaller order of magnitude than $n^+(\alpha, \epsilon)$ as $\epsilon \to 0$. The sample size required to control the overall error is modest especially in Cases 1 and 2. This is, however, no longer true for discrete populations as we shall see in Section 4.

The mathematics behind the tables. We shall write

$$\gamma_1 = \mu_3/\sigma^3, \qquad \gamma_2 = \mu_4/\sigma^4 - 3 \qquad \text{and}$$

$$H_2(t) = t^2 - 1, \qquad H_3(t) = t^3 - 3t, \qquad H_5(t) = t^5 - 10t^3 + 15t.$$

Case 1. Here (see for example Feller (1971), Chapter XVI)

$$P(t) = -\gamma_1 H_2(t)/6, \qquad Q(t) = -\gamma_2 H_3(t)/24 - \gamma_1^2 H_5(t)/72.$$

Therefore,

$$\beta_\alpha = \gamma_1 B(\alpha), \qquad \delta_\alpha = \gamma_1^2 D_1(\alpha) + \gamma_2 D_2(\alpha)$$

where

$$B(\alpha) = H_2(z_{\alpha/2})\phi(z_{\alpha/2})/6,$$
$$D_1(\alpha) = H_5(z_{\alpha/2})\phi(z_{\alpha/2})/36, \qquad D_2(\alpha) = H_3(z_{\alpha/2})\phi(z_{\alpha/2})/12.$$

Table 4 gives the values for $\alpha = 1, 5, 10\%$, and hence also the values for Case 1 in Table 1.

Table 4.

α	$z_{\alpha/2}$	$B(\alpha)$	$D_1(\alpha)$	$D_2(\alpha)$
1%	2.576	0.0136	− 0.0076	0.0113
5%	1.960	0.0277	− 0.0275	0.0080
10%	1.645	0.0293	− 0.0136	− 0.0042

The main reason why the bounds in Table 3 are simpler is that $H_3(\sqrt{3}) = 0$. It is left to the reader to verify Table 3 in the remaining cases.

Case 2. Here (Robinson (1978))

$$P(t) = -\gamma_1 H_2(t)(q-p)q^{-1/2}/6,$$
$$Q(t) = -((1-6pq)\gamma_2 - 6pq)H_3(t)q^{-1}/24 - (q-p)^2\gamma_1^2 H_5(t)q^{-1}/72.$$

Therefore,

$$\beta_\alpha = \gamma_1 B(\alpha)(q-p)q^{-1/2},$$
$$\delta_\alpha = \gamma_1^2 D_1(\alpha)(q-p)^2/q + ((1-6pq)\gamma_2 - 6pq)D_2(\alpha)/q.$$

Note that in this case the sample size n occurs even to the right in (2.2) and (2.3). Solving for n we get (2.4) and (2.5) respectively, provided $N > -a$ and $N > -b$. Here $c = -(\gamma_1^2 D_1(\alpha) + \gamma_2 D_2(\alpha))/\epsilon$, $a = 4|c| + b$, $b = \text{sign}(c)(2\gamma_2 + 6)D_2(\alpha)/\epsilon$ and $C = (\gamma_1 B(\alpha)/\epsilon)^2$.

Case 3. Here,

$$P(t) = \gamma_1(2t^2+1)/6,$$
$$Q(t) = -(t^3+t)/4 - \gamma_1^2(t^5+2t^3-3t)/18 + \gamma_2(t^3-3t)/12.$$

P is given by Hall (1983) and Abramovitch and Singh (1985). I have calculated Q using the method described in Bhattacharya and Ghosh (1978). So in this case

$$\beta_\alpha = \gamma_1 C(\alpha), \qquad \delta_\alpha = E_0(\alpha) + \gamma_1^2 E_1(\alpha) + \gamma_2 E_2(\alpha).$$

Here with $z = z_{\alpha/2}$

$$C(\alpha) = -(2z^2+1)\phi(z)/6, \qquad E_0(\alpha) = (z^3+z)\phi(z)/2,$$
$$E_1(\alpha) = (z^5+2z^3-3z)\phi(z)/9, \qquad E_2(\alpha) = -(z^3-3z)\phi(z)/6.$$

Table 5 gives the special values.

Table 5.

α	$C(\alpha)$	$E_0(\alpha)$	$E_1(\alpha)$	$E_2(\alpha)$
1%	− 0.0344	0.1421	0.2247	− 0.0226
5%	− 0.0846	0.2773	0.2474	− 0.0161
10%	− 0.1102	0.3143	0.1835	0.0083

Another possibility is to use the interval $\bar{x} \pm t_{\alpha/2}(n-1)s/n^{1/2}$ which is based on the t-distribution instead of the normal distribution. Which is best? Write T_n for the t-distribution with $n-1$ degrees of freedom. Then T_n has the same expansion as F_n but with $\gamma_1 = \gamma_2 = 0$. Therefore $t_{\alpha/2}(n-1) = z + O(n^{-1})$, and

$$F_n(t) = T_n(t) + n^{-1/2}P(t)\phi(t) + n^{-1}(Q(t) - Q_0(t))\phi(t) + o(n^{-1}),$$

and hence,

$$\tilde{\alpha}_n^+ = \alpha/2 - n^{-1/2}P(z)\phi(z) + O(n^{-1}),$$
$$\tilde{\alpha}_n = \alpha - n^{-1}2(Q(z) - Q_0(z))\phi(z) + o(n^{-1}).$$

Here $Q_0(t) = -(t^3+t)/4$, $z = z_{\alpha/2}$, and $\tilde{\alpha}_n$ and $\tilde{\alpha}_n^+$ are defined as α_n respectively α_n^+ but with $z_{\alpha/2}$ replaced by $t_{\alpha/2}(n-1)$. The one-sided bounds will therefore be the same for the two intervals, but the principal error term in the approximation of the overall confidence level will be smaller for the interval based on the t-distribution if $|Q(z)-Q_0(z)| < |Q(z)|$, that is if $Q(z) < Q_0(z)/2$. This is equivalent to

$$9(z^2+1) + \gamma_1^2 4(z^4+2z^2-3) - \gamma_2 6(z^2-3) > 0.$$

Note that $\gamma_2 \geq -2$, and therefore the above inequality is satisfied for all γ_1 and γ_2 if $(9/7)^{1/2} < z < 3^{1/2}$. That is, if α is between 8% and 26%.

The two-sided bounds are

$$n > |E_0(\alpha) + g_1^2 E_1(\alpha) + g_2 E_2(\alpha)|/\epsilon, \qquad n > |g_1^2 E_1(\alpha) + g_2 E_2(\alpha)|/\epsilon$$

for the normal interval and the t-interval, respectively.

Case 4. Let $U = n^{1/2}(\bar{x}-\mu)/\sigma(\mu)$, then $T/n^{1/2} = f(U/n^{1/2})$ where

$$f(\omega) = \omega\sigma(\mu)/\sigma(\mu + \sigma(\mu)\omega)$$

is defined in a neighbourhood of the origin, has two continuous derivatives and $f'(0) = 1$. Therefore there is a neighbourhood of the origin, $|\omega| < \delta$, in which f has an inverse satisfying

$$f^{-1}(\tau) = \tau + \sigma'(\mu)\tau^2 + (\sigma'(\mu)^2 + \sigma''(\mu)\sigma(\mu)/2)\tau^3 + o(\tau^3).$$

Hence if n is so large that $f(-\delta) < t/n^{1/2} < f(\delta)$, then $|U| < \delta n^{1/2}$ and $T < t$ if and only if $|U| < \delta n^{1/2}$ and $U \leq n^{1/2} f^{-1}(t/n^{1/2})$. Furthermore, the probability that $|U| \geq \delta n^{1/2}$ equals $o(1/n)$ and can therefore be neglected. This and the expansion for Case 1 yield

$$P(t) = -\gamma_1 H_2(t)/6 + \sigma'(\mu)t^2,$$
$$Q(t) = Q_1(t) + t^3(\gamma_1\sigma'(\mu)(t^2 - 3)/6 - \sigma'(\mu)^2(t^2 - 2)/2 + \sigma''(\mu)\sigma(\mu)/2)$$

where Q_1 is as Q in Case 1.

Therefore

$$\delta_\alpha = \gamma_1^2 D_1(\alpha) + \gamma_2 D_2(\alpha) + \gamma_1\sigma'(\mu)F_1(\alpha) + \sigma'(\mu)^2 F_2(\alpha) + \sigma(\mu)\sigma''(\mu)F_3(\alpha),$$
$$\beta_\alpha = \gamma_1 B(\alpha) + \sigma'(\mu)G(\alpha).$$

Here with $z = z_{\alpha/2}$

$$F_1(\alpha) = -\phi(z)z^3(z^2 - 3)/3, \qquad F_2(\alpha) = \phi(z)z^3(z^2 - 2),$$
$$F_3(\alpha) = -\phi(z)z^3, \qquad G(\alpha) = -\phi(z)z^2.$$

So in this case we get Table 6.

Table 6.

α	$G(\alpha)$	$F_1(\alpha)$	$F_2(\alpha)$	$F_3(\alpha)$
1%	− 0.0959	− 0.2994	1.1453	− 0.2471
5%	− 0.2245	− 0.1234	0.8104	− 0.4400
10%	− 0.2790	0.0450	0.3241	− 0.4590

The bounds can now be obtained from these expressions by replacing μ by $\bar{x}$ and γ_1 and γ_2 by g_1 and g_2, respectively.

3. A numerical illustration

Consider Case 1 when X_1 is exponentially distributed with expectation μ. In this case $\gamma_1 = 2$, $\gamma_2 = 6$ and $\sigma(\mu) = \mu$.

The interval (2.1) equals $\bar{x}/(1 \pm n^{-1/2}z_{\alpha/2})$. Here the right endpoint of the interval has to be replaced by ∞ if $n \leq z_{\alpha/2}^2$.

It is an amazing fact that $|\alpha_n - 5\%| < 1\%$ for *all* $n \geq 1$ when $\alpha = 5\%$ whereas $|\alpha_n^+ - 2.5\%| < 0.5\%$ only for $n \geq 110$, and $|\alpha_n^+ - 2.5\%| < 1\%$ only for $n \geq 24$.

Let n_E and n_E^+ denote the smallest sample sizes such that $|\alpha_n - \alpha| < \epsilon$ for all $n \geq n_E$ respectively $|\alpha_n^+ - \alpha/2| < \epsilon/2$ for all $n \geq n_E^+$, and let n_A and n_A^+ denote the smallest sample sizes recommended by the rule i.e. $n_A = [\delta_\alpha/\epsilon] + 1$ and $n_A^+ = [(2\beta_\alpha/\epsilon)^2] + 1$. Furthermore, let $\epsilon_E = \max\{|\alpha_n - \alpha|;\ n \geq n_E\}$, $\epsilon_A = \max\{|\alpha_n - \alpha|;\ n \geq n_A\}$, $\epsilon_E^+ = \max\{|\alpha_n - \alpha|;\ n \geq n_E^+\}$ and $\epsilon_A^+ = \max\{|\alpha_n - \alpha|;\ n \geq n_A^+\}$ denote the worst possible actual errors.

Table 7 gives a numerical illustration where α and ϵ are expressed in %.

Table 7.

α	ϵ	n_E	ϵ_E	n_A	ϵ_A	$\epsilon/2$	n_E^+	ϵ_E^+	n_A^+	ϵ_A^+
1	0.5	9	0.47	8	0.53	0.25	129	0.25	115	0.27
5	1	1	0.67	7	0.60	0.5	110	0.50	123	0.47
10	2	6	1.85	4	3.22	1	23	0.99	35	0.84

If we instead use the statistic in Case 4 we get the interval $\bar{x}(1 \pm n^{-1/2} z_{\alpha/2})$. In this case we get Table 8.

Table 8.

α	ϵ	n_A	ϵ_A	$\epsilon/2$	n_A^+
1	0.5	117	0.48	0.25	758
5	1	51	0.98	0.5	1144
10	2	17	1.86	1	486

Here and in Table 9 we use the same units as in the tables in Case 1.

If we, instead use the statistic in Case 3 we get the interval $\bar{x} \pm n^{-1/2} s z_{\alpha/2}$ and Table 9.

Table 9.

α	ϵ	n_A	$\epsilon/2$	n_A^+
1	0.5	181	0.25	758
5	1	118	0.5	1144
10	2	55	1	486

If we, in this case, instead use the interval based on the t-distribution, we get the two-sided bounds $n_A = 153, 90, 29$ which are smaller, but we have to know the value of $t_{\alpha/2}(n_A - 1)$.

Note that in all three cases the intervals have lengths that equal $2z_{\alpha/2}\mu n^{-1/2}+$ $O(n^{-1})$ a.s.

4. Discrete populations

What has been said above is no longer true when the population is not continuous but $X_1, \ldots, X_n$ take values in the set $\{a + kh;\ k = 0, 1, 2, \ldots\}$. Here h is the maximal number with this property. It can be argued that all distributions are discrete since our observations are measured in some smallest unit. It will be seen from the expansions below that a distribution can be considered as being continuous if h/σ is small compared to 1. Otherwise it is discrete.

We shall in this section, only consider Cases 1, 2 and 4. In Table 10 all bounds can be modified to hold for other epsilons via the formula $B(\alpha, \epsilon_2) = B(\alpha, \epsilon_1)(\epsilon_1/\epsilon_2)^2$.

Table 10. Bounds for sample size in Cases 1 and 4.

	α	ϵ	Overall error ϵ	Error $\epsilon/2$ in each tail
Case 1	1	0.5	$8(h/s)^2$	$(5.4\lvert g_1\rvert + 2.9h/s)^2$
	5	1	$34(h/s)^2$	$(5.5\lvert g_1\rvert + 5.8h/s)^2$
	10	2	$27(h/s)^2$	$(2.9\lvert g_1\rvert + 5.2h/s)^2$
Case 4	1	0.5	$8(h/\sigma(\bar{x}))^2$	$(\lvert 5.4g_1 - 38.4\sigma'(\bar{x})\rvert + 2.9h/\sigma(\bar{x}))^2$
	5	1	$34(h/\sigma(\bar{x}))^2$	$(\lvert 5.5g_1 - 44.9\sigma'(\bar{x})\rvert + 5.8h/\sigma(\bar{x}))^2$
	10	2	$27(h/\sigma(\bar{x}))^2$	$(\lvert 2.9g_1 - 27.9\sigma'(\bar{x})\rvert + 5.2h/\sigma(\bar{x}))^2$

The corresponding bounds for Case 2 are

$$\min(n, N-n) > \frac{2A^2}{1+\sqrt{1-4\dfrac{A^2}{N}}} \tag{4.1}$$

respectively

$$\min(n, N-n) > \frac{2(A+B)^2}{1+\dfrac{4B(A+B)}{N}+\sqrt{1+4\dfrac{B^2-A^2}{N}}}. \tag{4.2}$$

Here A^2 is the bound for the overall error in Case 1, $A > 0$, and $(A+B)^2$ is the bound in the last column in Case 1. We have assumed that N is so large that the expressions under the squareroot signs are positive.

The bound

$$n\bar{x}(1-\bar{x}) > 10,$$

which sometimes is used in connection with the binomial distribution, thus correspond to our two-sided bound when $\alpha = 1\%$, $\epsilon = 0.5\%$ or $\alpha = 5\%$, $\epsilon = 1.8\%$ or $\alpha = 10\%$, $\epsilon = 3.2\%$ for example. The corresponding bound for the hypergeometric distribution is

$$n\bar{x}(1-\bar{x}) > \frac{20}{1+\sqrt{1-\dfrac{40}{N\bar{x}(1-\bar{x})}}}.$$

Cases 1 *and* 2. The expansion (1.3) can in Cases 1 and 2 be modified to hold even for discrete distributions. We shall not need the Q-term to begin with. The modified expansion is

$$F_n(t) = \Phi(s) + n^{-1/2}P(s)\phi(s) + O(1/n). \tag{4.3}$$

It is valid when $t = (na + kh - n\mu)/(\tau n^{1/2})$ for some integer k. Here $\tau = \sigma$ in Case 1, $\tau = \sigma(qN/(N-1))^{1/2}$ in Case 2, and $s = t + h/(2\tau n^{1/2})$.

Put $\eta = h/(\tau n^{1/2})$ and $z = z_{\alpha/2}$. Choose $t_j = (na + k_j h - n\mu)/(\tau n^{1/2})$ and real numbers $-1/2 \le \omega_1 \le 1/2$, $-1/2 \le \omega_2 \le 1/2$ such that

$$z = t_1 + \left(\frac{1}{2} - \omega_1\right)\eta, \qquad -z = t_2 + \left(\frac{1}{2} - \omega_2\right)\eta.$$

Then,

$$\alpha_n^+ = 1 - F_n(t_1), \qquad \alpha_n = 1 - F_n(t_1) + F_n(t_2) \quad \text{and}$$

$$s_1 = t_1 + \eta/2 = z + \omega_1\eta, \qquad s_2 = t_2 + \eta/2 = -z + \omega_2\eta.$$

Taylor expansions of the expression to the right in (4.3) around the points $s_1 = z$ and $s_2 = -z$ therefore yield

$$\alpha_n^+ = \alpha/2 + n^{-1/2}(\beta_\alpha - \omega_1\phi(z)h/\tau) + O(n^{-1}),$$
$$\alpha_n = \alpha + n^{-1/2}(\omega_2 - \omega_1)\phi(z)h/\tau + O(n^{-1}).$$

Note that ω_1 and ω_2 vary with μ and σ, and that we cannot estimate this variation with sufficient precision. We shall therefore maximize the principal error term in order to be on the safe side. The maximum is attained in the one-sided case when ω_1 has the same sign as $-\beta_\alpha$ and modulus $1/2$, and in the two-sided case when ω_1 and ω_2 have modulus $1/2$ and opposite signs. In this way we get the inequalities

$$n > \left(\frac{\phi(z_{\alpha/2})h}{\tau\epsilon}\right)^2 \tag{4.4}$$

and

$$n > \left(\frac{\phi(z_{\alpha/2})h}{2\tau\epsilon} + \frac{|\beta_\alpha|}{\epsilon}\right)^2 \tag{4.5}$$

corresponding to (2.2) and (2.3), respectively. Thus if B_d denotes the two-sided bound in the discrete case corresponding to ϵ, and B_{d+} and B_{c+} the one-sided

bounds corresponding to $\epsilon/2$ in the discrete and continuous case, respectively. Then in Case 1: $B_{d+}^{1/2} = B_{c+}^{1/2} + B_d^{1/2}$.

In Case 2, (4.4) and (4.5) are equivalent to (4.1) and (4.2) respectively, provided $N > 4A^2$. Here

$$A = \frac{\phi(z_{\alpha/2})h}{\epsilon\sigma}\frac{N-1}{N} \quad \text{and} \quad B = \frac{|\gamma_1 B(\alpha)|}{\epsilon}.$$

The bounds will change very little by replacing the quotient $(N-1)/N$ by 1.

Note that, if h is close to 0, then the one-sided bound is close to the corresponding bound in the continuous case. This is as it should be, but the two-sided bound degenerates to $n > 0$. This is so because the principal error term in the approximation $\alpha_n \approx \alpha$ comes from the next term in the expansion (4.3). The next term is $n^{-1}Q_h(s)\phi(s)$ where

$$Q_h(s) = Q(s) - (s/24)(h/\tau)^2.$$

Therefore,

$$\begin{aligned}\alpha_n &= \alpha + (\omega_1 - \omega_2)\eta\phi(\zeta) - 2Q_h(z)\phi(z)/n - (\omega_1^2 + \omega_2^2)\eta^2\phi'(z) \\ &\quad - (\omega_1 + \omega_2)\eta n^{-1/2}(P'(z)\phi(z) + P(z)\phi'(z)) + o(n^{-1}) \\ &= \alpha + (\omega_2 - \omega_1)\eta\phi(z) + n^{-1}\delta_\alpha + O(\eta n^{-1/2}) + o(n^{-1}).\end{aligned}$$

The worst case gives the requirement $\eta\phi(z) + n^{-1}|\delta_\alpha| < \epsilon$, which is equivalent to

$$n > \left(\frac{\phi(z_{\alpha/2})h}{2\tau\epsilon} + \sqrt{\left(\frac{\phi(z_{\alpha/2})h}{2\tau\epsilon}\right)^2 + \frac{|\delta_\alpha|}{\epsilon}}\right)^2. \tag{4.6}$$

The bounds (4.6) and (4.5) can thus in Cases 1 and 2 serve as bounds in both the discrete and continuous case provided we agree that $h = 0$ in the continuous case.

Case 4. Modifying the argument leading to the expansion for the distribution of the Case 4 statistic in the continuous case, we get in the discrete Case 4

$$\begin{aligned}\alpha_n &= \alpha + n^{-1/2}(\omega_2 - \omega_1)\phi(z)h/\sigma + O(n^{-1}), \\ \alpha_n^+ &= \alpha/2 - n^{-1/2}\phi(z)(P(z) + \omega_1 h/\sigma) + O(n^{-1})\end{aligned}$$

where $|\omega_i| \le 1/2$, and where P is as in the continuous Case 4. The worst case then gives the bounds

$$n > \left(\frac{\phi(z_{\alpha/2})h}{\sigma(\bar{x})\epsilon}\right)^2,$$

$$n > \left(|g_1 B(\alpha) + \sigma'(\bar{x})G(\alpha)| + \frac{\phi(z_{\alpha/2})h}{2\sigma(\bar{x})}\right)^2 \Big/ \epsilon^2.$$

Example. Let us as a final example compare the confidence intervals Cases 1 and 4 will give for the success probability in a sequence of Bernoulli trials. In Case 1 we get

$$\left(\bar{x} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n} + \left(\frac{z_{\alpha/2}}{2n}\right)^2}\right) \Big/ (1 + z_{\alpha/2}^2/n)$$

and in Case 4

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}.$$

The two-sided bound will in both cases be

$$n\bar{x}(1-\bar{x}) > \left(\frac{\phi(z_{\alpha/2})}{\epsilon}\right)^2$$

whereas the one-sided bounds are

$$n\bar{x}(1-\bar{x}) > \left(\frac{(|1-2\bar{x}||z_{\alpha/2}^2 - 1| + 3)\phi(z_{\alpha/2})}{6\epsilon}\right)^2$$

for the first interval and

$$n\bar{x}(1-\bar{x}) > \left(\frac{(|1-2\bar{x}|(2z_{\alpha/2}^2 + 1) + 3)\phi(z_{\alpha/2})}{6\epsilon}\right)^2$$

for the second. The former bound is smaller than the latter except when $\bar{x} = 1/2$ in which case they are equal.

References

Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap, *Ann. Statist.*, **13**, 116–132.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of formal Edgeworth expansion, *Ann. Statist.*, **6**, 434–451.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed., Wiley, New York.

Dalén, J. (1986). Sampling from finite populations: Actual coverage probabilities for confidence intervals on the population mean, *Journal of Official Statistics*, **2**, 13–24.

Feller, W. (1971). *An Introduction to Probability Theory*, Vol. 2, 2nd ed., Wiley, New York.

Hall, P. (1983). Inverting an Edgeworth expansion, *Ann. Statist.*, **11**, 569–576.

Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations, *J. Amer. Statist. Assoc.*, **73**, 536–544.

Robinson, J. (1978). An asymptotic expansion for samples from a finite population, *Ann. Statist.*, **6**, 1005–1011.