Confidence Intervals for a Mean and a Proportion in the Bounded Case

George S. Fishman

Technical Report No. UNC/ORSA/TR-86/19
November 1986

UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL
Confidence Intervals for a Mean and a Proportion in the Bounded Case

George S. Fishman

Technical Report No. UNC/ORSA/TR-86/19

November 1986

Curriculum in Operations Research and Systems Analysis
University of North Carolina
Chapel Hill, North Carolina

This research was supported by the Air Force Office of Scientific Research under grant AFOSR-84-0140. Reproduction in whole or part is permitted for any purpose of the United States Government.
This paper describes a $100 \times (1-\alpha)$ confidence interval for the mean of a bounded random variable which is shorter than the interval that Chebyshev's inequality induces for small $\alpha$ and which avoids the error of approximation that assuming normality induces. The paper also presents an analogous development for deriving a $100 \times (1-\alpha)$ confidence interval for a proportion.
Abstract

This paper describes a \(100\times(1-\alpha)\) confidence interval for the mean of a bounded random variable which is shorter than the interval that Chebyshev's inequality induces for small \(\alpha\) and which avoids the error of approximation that assuming normality induces. The paper also presents an analogous development for deriving a \(100\times(1-\alpha)\) confidence interval for a proportion.

Key words: Confidence interval, Proportion.

Acknowledgement: I am grateful to Tien Yi Shaw for his assistance in preparing this paper.
Introduction

Let $X_1, \ldots, X_n$ be independent identically distributed random variables with $\mu = E X_i$, $pr(0 \leq X_i \leq 1) = 1$ and $\bar{X}_n = (X_1 + \ldots + X_n)/n$. This paper describes a method for deriving a $100(1-\alpha)$ interval estimate of $\mu$ for finite $n$ based on the probability inequality (Hoeffding 1963, Thm. 1, (2.1))

$$pr(\bar{X}_n - \mu \pm \epsilon) \leq e^{nf(\epsilon, \mu)} \quad (1)$$

where

$$f(\epsilon, \mu) = (\epsilon + \mu)[ln \mu - ln(\mu + \epsilon)] + (1 - \epsilon - \mu)[ln(1 - \mu) - ln(1 - \mu - \epsilon)] \quad \epsilon < 1 - \mu \quad (2)$$

and

$$\lim_{\epsilon \to 1 - \mu} f(\epsilon, \mu) = ln \mu.$$

To put our results in perspective, we first review the derivation of two commonly encountered confidence intervals. Let

$$k(x, \beta, m) = \left\{ x + \beta^2/2m + \beta [ \beta^2/4m + x(1-x)/m ]^{1/2} \right\} / (1 + \beta^2/m)$$

$$0 \leq x \leq 1, \quad -\infty < \beta < \infty, \quad m = 1, 2, \ldots. \quad (3)$$

Then the interval $(k(\bar{X}_n, - \phi^{-1}(1/2), n), k(\bar{X}_n, \phi^{-1}(1/2), n))$ covers $\mu$ with confidence coefficient $> 1 - \alpha$, as a consequence of Chebyshev's inequality and the observation that $var X_i = E(X_i - \mu)^2 = E[X_i(X_i - \mu)] \leq (1 - \mu)$. Moreover, as a result of the central limit theorem, the interval $(k(\bar{X}_n, - \phi^{-1}(1 - \alpha/2), n), k(\bar{X}_n, \phi^{-1}(1 - \alpha/2), n))$ asymptotically (as $n \to \infty$) covers $\mu$ with confidence coefficient $\geq 1 - \alpha$, where

$$\phi^{-1}(\alpha) = \{ y : (2\pi)^{-1/2} \int_{-\infty}^{y} e^{-z^2/2} \, dz = \alpha \}.$$
Although the Chebyshev interval holds for all \( n \), the width of the resulting interval is considerably larger than that for the asymptotic normal one. For example \( \alpha^{-\frac{1}{2}} \phi^{-1}(1-\alpha/2) = 2.28 \) for \( \alpha=.05 \). However, because of the nonuniform convergence of \( (\bar{X}_n - \mu)/[\mu(1-\mu)/n]^{\frac{1}{2}} \), using the normal confidence interval obliges one to account for the inevitable error of approximation for finite \( n \). This error makes difficult an assessment of whether or not the associated confidence coefficient truly exceeds \( 1-\alpha \), and can be especially bothersome in a Monte Carlo sampling experiment where the problem dictates the maximal interval width and the minimal acceptable confidence level. Even less appealing are interval estimates of the form \( (\bar{X}_n - \beta(S^2/n)^{\frac{1}{2}}, \bar{X}_n + \beta(S^2/n)^{\frac{1}{2}}) \) where

\[
S^2_n = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]

and \( \beta = \alpha^{-\frac{1}{2}} \) and \( \beta = \phi^{-1}(1-\alpha/2) \) for the Chebyshev and normal cases respectively. Although intended to shorten the intervals by using the additional information in \( S^2_n \leq \bar{X}_n(1-\bar{X}_n) \) the substitution of \( S^2_n \) for \( \text{var} X_i \) induces an additional error of approximation in assessing whether or not the resulting confidence coefficient exceeds \( 1-\alpha \).

Hoeffding (1963) derived the probability inequality (1) for all bounded \( X_i \). Previously Okamoto had derived (2) for \( X_i \) with the Bernoulli distribution, \( \text{pr}(X_i=0) = 1-\mu \) and \( \text{pr}(X_i=1) = \mu \), a result implicit in Chernoff (1952, Thm. 1 and Ex. 5). Theorem 1
provides the basis for constructing a confidence interval for $\mu$ based on Hoeffding's theorem.

**Theorem 1.** Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mu = \text{EX}_1$, $\text{pr}(0 \leq X_i \leq 1) = 1$ and $\lambda = \max \left[ \text{pr}(X_i = 0), \text{pr}(X_i = 1) \right]$. Then for $n \geq \ln(a/2)/\ln\lambda$, $(\bar{\psi}_1(\bar{x}_n, a/2), \bar{\psi}_2(\bar{x}_n, a/2))$ covers $\mu$ with probability $> 1 - \alpha$ where $\bar{\psi}_1(\bar{x}_n, a/2) \leq \bar{x}_n \leq \bar{\psi}_2(\bar{x}_n, a/2)$ are the solutions to

$$f(\bar{x}_n - \psi, \psi) = \frac{1}{n} \ln(a/2). \quad (4)$$

**Proof.** Observe that

$$df(\epsilon, \mu)/d\epsilon = \ln[\mu(1-\mu-\epsilon)/(\mu+\epsilon)(1-\mu)] < 0 \quad 0 \leq \epsilon < 1 - \mu,$$

$$f(0, \mu) = 0$$

and recall that

$$\lim_{\epsilon \to 1 - \mu} f(\epsilon, \mu) = \ln \mu < 0.$$ 

Therefore, $\text{enf}(\epsilon, \mu)$ is monotone decreasing in $\epsilon$ with maximum 1 and minimum $\mu^n$.

Consider the equal tail probability case and let

$$\epsilon(\mu, a/2) = \{\epsilon: \text{enf}(\epsilon, \mu) = a/2\}$$

if $\mu^n \leq a/2$

$$= 1 - \mu$$

if $\mu^n > a/2$.

Then

$$\text{pr}[\bar{x}_n \geq \epsilon(\mu, a/2)] \leq a/2 \quad (5)$$
where
\[ h(\mu, \alpha/2) = \mu + \varepsilon(\mu, \alpha/2). \]

For \( \mu^n > \alpha/2 \), \( \Pr(\bar{X}_n \geq 1) \leq \lambda^n \leq \alpha/2 \). For \( \mu^n \leq \alpha/2 \), we want to find the set of all \( \mu \)'s satisfying (5). From \( \epsilon(\varepsilon, \mu) = \alpha/2 \),

\[ d\epsilon(\mu, \alpha/2)/d\mu = -\{1+\varepsilon/\mu(1-\mu) \ln[\mu(1-\varepsilon)/(1-\mu)(\mu+\varepsilon)] \} \]  

so that
\[ dh(\mu, \alpha/2)/d\mu > 0, \]

implying that \( h \) is monotone increasing in \( \mu \). Therefore, the set of \( \mu \)'s of interest is \( \{\mu: 0 < \mu \leq \Psi_1(\bar{X}_n, \alpha/2)\} \) where
\[ \Psi_1(x, \alpha/2) = \{\Psi: \Psi + \varepsilon(\Psi, \alpha/2) = x\}, \]

which is precisely the solution to (4) in the interval \([0, \bar{X}_n]\). Consequently,
\[ \Pr(\Psi_1(\bar{X}_n, \alpha/2) \leq \mu) \leq \alpha/2 \]

so that
\[ \Pr(\Psi_1(\bar{X}_n, \alpha/2) < \mu) > 1 - \alpha/2, \]

as required.

The upper bound \( \Psi_2(\bar{X}_n, \alpha/2) \) follows analogously, using
\[ \Pr(-\bar{X}_n + \mu \geq \varepsilon) \leq \epsilon(\varepsilon, 1-\mu). \]
Observe that if $X_i$ has a continuous distribution, $\lambda = 0$ and Theorem 1 holds for all sample sizes $n$. If $\Pr(a \leq X \leq b) = 1$, then Theorem 1 holds with $100\times(1-\alpha)$ confidence interval $((b-a) \Psi_1((X_n-a)/(b-a),\alpha/2) + a, (b-a) \Psi_2((X_n-a)/(b-a),\alpha/2) + a)$. Although for Bernoulli data and small $n$, one can compute an exact confidence interval for $\mu$, as in Blyth and Still (1983), this option loses its appeal as $n$ increases and the potential for numerical error grows. Table 1 shows the lower bounds on $n$ for $\alpha = .01$ and .05.

Using the dominant term (as $n \to \infty$) of the Taylor series of $f(x; \psi, \psi)$, one can readily show that as $n$ increases

$$
\Psi_2(\bar{X}_n, \alpha/2) - \Psi_1(\bar{X}_n, \alpha/2) = 2[2 \ln(2/\alpha) \bar{X}_n(1-\bar{X}_n)/n]^{1/2}.
$$

To order $n^{-1/2}$, the Chebyshev and normal intervals have widths $2[\alpha^{-1} \bar{X}_n(1-\bar{X}_n)/n]^{1/2}$ and $2\Phi^{-1}(1-\alpha/2)[\bar{X}_n(1-\bar{X}_n)/n]^{1/2}$ respectively. Table 2 compares these widths for $\alpha = .01$ and .05.

Confidence Interval for a Proportion

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote i.i.d. random vectors with $\mu_X = EX_1$, $\mu_Y = EY_1$, $\Pr(0 \leq X_1 \leq 1) = 1$, $\Pr(0 \leq Y_1 \leq 1) = 1$, $\Pr(Y_1 \leq X_1) = 1$, $\phi = \mu_Y/\mu_X$, $\bar{X}_n = (X_1 + \ldots + X_n)/n$ and $\bar{Y}_n = (Y_1 + \ldots + Y_n)/n$. Also, let
\[ r(x, y, \beta, m) = \left\{ xy + \beta^2/m + \beta \left[ \beta^2/4m^2 + y(x-y)/m \right]^{1/2} \right\} / (x^2 + \beta^2/m) \]

\[ 0 \leq y \leq 1, \quad -\infty < \beta < \infty, \quad m = 1, 2, \ldots \]

(8)

Then \( r(\bar{x}_n, \bar{y}_n, -\beta, n), r(\bar{x}_n, \bar{y}_n, \beta, n) \) with \( \beta = \alpha^{-1/2} \) covers \( \phi \) with confidence coefficient \( > 1-\alpha \), and with \( \beta = \phi^{-1}(1-\alpha/2) \) asymptotically \( (n \to \infty) \) covers \( \phi \) with confidence coefficient \( 1-\alpha \). These results again follow from Chebyshev's inequality, the central limit theorem and the observation that

\[ \text{var}(Y_i - \phi X_i) = \text{var}(Y_i - \phi X_i + \phi) \leq \phi(1-\phi). \]

Again, one can derive a 100\( \times (1-\alpha) \) confidence interval, shorter than the one that Chebyshev's inequality offers for small \( \alpha \) and that avoids the error of approximation that assuming normality induces. Let

\[ W_i = Y_i - \phi X_i + \phi \]

so that \( \phi = E W_i \) and \( \Pr(0 \leq W_i \leq 1) = 1 \). Then for \( \bar{W}_n = (W_1 + \ldots + W_n)/n \), (1) applies in the form

\[ \Pr(Y_i - \phi X_i \geq \varepsilon) = \Pr(W_i - \phi \varepsilon) \leq e^{-\text{nf}(\varepsilon, \phi)}. \]

(9)

This establishes the basis for Theorem 2.

**Theorem 2.** Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be i.i.d. random vectors with \( \mu_X = E X_i \), \( \mu_Y = E Y_i \), \( \Pr(0 \leq X_i \leq 1) = 1 \), \( \Pr(0 \leq Y_i \leq 1) = 1 \), \( \Pr(Y_i \leq X_i) = 1 \), \( \lambda = \max \Pr(Y_i = 0) \), \( \Pr(Y_i = 1) \), \( \bar{x}_n = (X_1 + \ldots + X_n)/n \), \( \bar{y}_n = (Y_1 + \ldots + Y_n)/n \) and \( \phi = \mu_Y / \mu_X \). Then for \( n \geq 1n(\alpha/2)/1n \lambda \), \((Y_1(\bar{x}_n, \bar{y}_n, \alpha/2), Y_2(\bar{x}_n, \bar{y}_n, \alpha/2))\) covers \( \phi \) with probability \( > 1-\alpha \)
where \( Y_1(x, y, \theta) \leq y/x \leq Y_2(x, y, \theta) \), are the solutions of

\[
f(y - y/x, y) = \frac{1}{n} \ln \theta \quad \text{for} \quad 0 \leq y \leq x \leq 1, \quad (10)
\]

\( f \) being defined in (2).

**Proof.** Let

\[
c(\phi, a/2) = \{ c : f(c, \phi) = \frac{1}{n} \ln(a/2) \} \quad \text{if} \quad \phi^n \leq a/2
\]

\[
= 1 - \phi \quad \text{if} \quad \phi^n > a/2.
\]

Then

\[
pr[\tilde{Y}_n \geq g(\phi, a/2, \tilde{X}_n)] \leq a/2 \quad (11)
\]

where

\[
g(\phi, a/2, x) = \phi x + \varepsilon(\phi, a/2).
\]

For \( \phi^n > a/2 \),

\[
pr(\tilde{Y}_n - \phi \tilde{X}_n \geq 1 - \phi) = pr(\tilde{Y}_n = 1, \tilde{X}_n = 1)
\]

\[
= pr(\tilde{X}_n = 1 | \tilde{Y}_n = 1) \cdot pr(\tilde{Y}_n = 1)
\]

\[
= pr(\tilde{Y}_n = 1)
\]

\[
= \lambda^n \leq a/2.
\]

For \( \phi^n \leq a/2 \), we want to find the set of all \( \phi \)'s satisfying (11). Observe that

\[
dg(\phi, a/2, x)/d\phi = x + \psi \varepsilon(\phi, a/2)/\partial \phi
\]

where (6) gives \( \psi \varepsilon(\phi, a/2)/\partial \phi \). Using the inequalities \( z/(1+z) < \ln(1+z) < z \) for \( z > -1 \) and \( z = 0 \), one has
Since $\epsilon > 0$, so that

$$-\phi / (1-\epsilon) \geq (\epsilon - \tilde{Y}_n) \epsilon / (1-\epsilon) (\tilde{X}_n - \tilde{Y}_n + \epsilon) > (\epsilon - \tilde{Y}_n) / (1-\epsilon),$$

and finally

$$3g(0, a/2, \tilde{X}_n) / \phi \geq \tilde{X}_n + (\epsilon - \tilde{Y}_n) / (1-\epsilon) = [\tilde{X} - \tilde{Y}_n + \epsilon (1-\tilde{X}_n)] / (1-\epsilon) > 0.$$ 

Therefore, the set of $\phi$'s of interest is $\{ \phi: 0 < \phi \leq Y_1(\tilde{X}_n, \tilde{Y}_n, a/2) \}$
where

$$Y_1(x, y, \phi) = \{ \psi: \psi x + \epsilon (\psi, a/2) = y, 0 \leq y \leq 1, 0 < \phi < 1 \},$$

which is precisely the solution to (10). Consequently,

$$Pr[Y_1(\tilde{X}_n, \tilde{Y}_n, a/2) > \phi] \leq a/2$$

so that

$$Pr[\phi > Y_1(\tilde{X}_n, \tilde{Y}_n, a/2)] > 1 - a/2,$$

as required. The upper bound $Y_2(\tilde{X}_n, \tilde{Y}_n, a/2)$ follows analogously using

$$Pr(\tilde{w}_n + \psi \geq \epsilon) \leq \exp(\epsilon, 1-\phi).$$
References


Table 1

\[ n_0 = \min \{ n : \lambda^n \leq \alpha/2 \} \]

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \lambda )</th>
<th>( n_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.01</td>
<td>.90</td>
<td>51</td>
</tr>
<tr>
<td>.99</td>
<td>528</td>
<td></td>
</tr>
<tr>
<td>.999</td>
<td>5296</td>
<td></td>
</tr>
<tr>
<td>.9999</td>
<td>52981</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \lambda )</th>
<th>( n_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.90</td>
<td>36</td>
</tr>
<tr>
<td>.99</td>
<td>368</td>
<td></td>
</tr>
<tr>
<td>.999</td>
<td>3688</td>
<td></td>
</tr>
<tr>
<td>.9999</td>
<td>36887</td>
<td></td>
</tr>
</tbody>
</table>

Table 2

Comparison of Interval Widths

\[
\begin{align*}
\alpha & & \alpha^{-1/2}/\phi^{-1}(1-\alpha/2) & & \alpha^{-1/2}/[2\ln(2/\alpha)]^{1/2} & & [2\ln(2/\alpha)]^{1/2}/\phi^{-1}(1-\alpha/2) \\
.01 & & 3.88 & & 3.07 & & 1.26 \\
.05 & & 2.28 & & 1.65 & & 1.39
\end{align*}
\]