Some Theory for the Stringer Bound of Auditing Practice

By

Peter J. Bickel
Department of Statistics
University of California
Berkeley, California

Technical Report No. 272
October 1990

Research partially supported by ONR Contract N00014-89-J-1563
NSA Grant MDA904-89-H-2045.

Department of Statistics
University of California
Berkeley, California

Some Theory for the Stringer Bound of Auditing Practice

Peter J. Bickel

Department of Statistics, University of California, Berkeley, California 94720

Summary

Accounting practice calls for nonparametric upper confidence bounds on the total error amount in accounting populations. Dollar unit sampling and the assumption that actual value never exceeds book value lead to the problem of setting nonparametric upper confidence bound on the mean of a population taking values between 0 and 1 on the basis of a sample from that population. The usual Gaussian asymptotic theory bounds are unsatisfactory since, though samples are large, there are few informative (nonzero) observations. An ad hoc bound the so called Stringer bound, has been found to be conservative and is widely used in accounting practice but its theoretical properties are essentially unknown. We give some weak fixed sample support to the bound's conservativeness and show that asymptotically it is essentially always too big. In addition we discuss a number of bounds which can be shown to be conservative and propose a simple new procedure which, initial simulations suggest, shares the conservatism of the Stringer bound for small numbers of nonzero observations and behaves like the asymptotically correct Gaussian based bound for larger numbers of nonzero observations.

Sommaire:

(French) La pratique de la comptabilité nécéssite des borgnes de confiance pour la somme totale des erreurs dans des populations de comptes. L'échantillonage ''Dollar unit'' mène au problème de mettre une borgne de

Research partially supported by ONR Contract N00014-89-J-1563, NSA Grant MDA904-89-H-2045

confiance nonparamétrique sur l'espérance d'une variable aléatoire prenant des valeurs entre 0 et 1. Les borgnes basées sur la thèorie asymptotique Gaussienne sont pas bonnes puisque bien que les échantillons sont grands la grande majorité des valeurs est noninformative (zero). Une borgne "ad hoc" appellée celle de Stringer est en grand emploi dans la profession des comptables puisque elle est conservative en pratique. Mais ses propriétés théoriques sont inconnues. Nous établissons quelques propriétés de la borgne surtout qu'elle est en effet toujours trop grande. Nous discutons aussi quelques borgnes qu'on peut démontrer sont consérvatives et aussi une nouvelle borgne qui ressemble Stringer quand le nombre d'observations positives est petit et ressemble la borgne Gaussienne quand le nombre est plus grand. Quelques simulations supportent notre candidat.

1 Introduction

The Stringer bound is a widely used nominal $100(1-\alpha)\%$ upper confidence bound for the total error amount in accounting populations when dollar unit sampling is employed. The bound has been found to be conservative in practice, often excessively so but nothing seems known of its theory. In this paper we partly remedy this lack and also discuss a number of alternative bounds. An excellent presentation of statistical issues in auditing and of the Stringer bound and other statistical techniques of auditing may be found in the N.R.C. report "Statistical Models and Analysis in Auditing" (1988) reprinted in Statistical Science (1989). We use this report as the basis of our presentation.

In auditing we are given a population $\{y_1, \ldots, y_N\}$ of "book values of items". From this population, by some random mechanism, n items labelled $j_1, \ldots, j_n$ are selected for audit. Let x_j denote the audited value of item j. Our observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are the audit and book values of the selected items. The ultimate goal is to set upper (or lower) confidence bounds on the population error $\Delta \equiv \Sigma_{j=1}^{N}(y_j - x_j)$. A company audit would typically result in an upper bound while the I.R.S. may be more interested in a lower bound.

One of the most popular schema for drawing samples is so called dollar unit sampling more commonly known as sampling proportional to size without replacement. That is, a first item is selected with probability y_j / Y of getting item j where $Y \equiv \Sigma_{i=1}^{N} y_i$. A second item is selected from the remainder with probability again proportional to the book value of the item selected etc. Since N is large in the situations considered it is plausible to approximate this scheme by sampling proportional to size with replacement which leads to (X_i, Y_i) being i.i.d. with

$$P\,[(X_i, Y_i) = (x_j, y_j)] \;=\; \frac{y_j}{Y}, \quad 1 \le j \le N .$$

If we assume nothing further about the population the UMVU estimate of Δ is just,

$$\hat{\Delta} \;=\; Y\, n^{-1} \sum_{i=1}^{n} T_i \tag{1.1}$$

where T_i the ''taint'' of the i-th selected item is given by, $T_i \equiv (Y_i - X_i)/Y_i$.

It is usually assumed and often valid in accounts receivable populations that $0 \le T_i \le 1$; only overstatements are possible with maximum error the book amount. In this form the problem of setting confidence bounds on Δ reduces to that of setting confidence bounds on $\mu \equiv E\,(T_1)$ based on an i.i.d. sample $T_1, \ldots, T_n$ when we know $0 \le T_i \le 1$. Standard normal and t approximations of course apply but have been found to be poor in practice. The reason seems to be that, as might be expected, most of the T_i are 0. The distribution of T is highly skewed and the number of observations available for estimating $E\,(T \mid T > 0)$, the crucial factor in $E\,(T)$ is small, sometimes 0. To deal with this problem the following upper confidence bound credited to Stringer is used. If $M \equiv$ number of non zero T_i let $0 < z_M \le \cdots \le z_1$ be the ordered non zero T_i. Let $p\,(j; 1-\alpha) \equiv 1-\alpha$ exact upper confidence bound for p when $X \sim bin\,(n, p)$ and $X = j$. Thus, $p\,(j; 1-\alpha)$ is the unique solution of

$$\sum_{k=j+1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \;=\; 1-\alpha . \tag{1.2}$$

The Stringer bound (for the mean taint μ) is,

$$\bar{\mu}_{ST} \;\equiv\; p\,(0; 1-\alpha) \;+\; \sum_{j=1}^{m} [\,p\,(j; 1-\alpha) - p\,(j-1; 1-\alpha)\,]\, z_j \tag{1.3}$$

Evidently,

$$P\,[\bar{\mu}_{ST} \geq \mu \,|\, M = 0] \;\geq\; P\,[\bar{\mu}_{ST} \geq P\,[T > 0]\,|\,M = 0] \;=\; 1 - \alpha. \tag{1.4}$$

In section 2 we motivate the Stringer bound and show more generally that,

$$\begin{aligned} P\,[\bar{\mu}_{ST} \geq \mu] \;&\geq\; (1-\alpha)^{n+1} \text{ for } n \geq 2 \\ &\geq\; (1-\alpha) \text{ for } n = 0, 1 \quad \text{or} \quad \pi = 0\,. \end{aligned} \tag{1.5}$$

On the other hand we show in section 2 that, as $n \to \infty$,

$$\underline{LIM}_n\, P\,[\bar{\mu}_{ST} \geq \mu] \;\geq\; 1 - \alpha \tag{1.6}$$

with strict inequality unless the distribution of T concentrates on at most two points other than 0. But it is also easy to see that the bound is conservative for distributions concentrating on at most one point other than 0. All this evidence suggests the bound is always conservative but we have not been able to show this.

In section 3 we briefly discuss some genuinely and approximately conservative alternatives to $\bar{\mu}_{ST}$ as well as other choices for an upper confidence bound.

2 The Stringer bound

Assume $0 \leq T \leq 1$.

To motivate the Stringer bound let $\pi \equiv P\,[T > 0]$, $\mu \equiv E\,(T)$, G be the conditional distribution of T, given $T > 0$. Then,

$$\mu \;=\; \pi \int_0^1 t\, dG\,(t) \tag{2.1}$$

and for $M, z_1, \ldots, z_M$ fixed,

$$\begin{aligned} &= \pi \sum_{j=0}^{M} \int_{(z_{j+1}, z_j]} t\, dG\,(t) \\ &\leq \pi \sum_{j=0}^{M} z_j\, (G\,(z_j) - G\,(z_{j+1}))\} \end{aligned}$$

where $z_0 \equiv 1$, $z_{M+1} = 0$. Then, by Abel partial summation,

$$\mu \;\leq\; \pi\{\sum_{j=0}^{M-1} (z_j - z_{j+1})(1 - G\,(z_{j+1})) + z_M\}\,. \tag{2.2}$$

Now, by Abel again,

$$\bar{\mu}_{ST} = \sum_{j=0}^{M-1} (z_j - z_{j+1}) p\,(j, 1-\alpha) + z_M\, p\,(M, 1-\alpha)\,. \tag{2.3}$$

Let $U_1, \ldots, U_n$ be i.i.d. uniform $(0,1)$ and $U_{(1)} < \cdots < U_{(n)}$ be the corresponding orders statistics. Then if G is continuous, it is easily seen that,

$$\mathbf{L}(1 - G\,(z_1), \ldots, 1 - G\,(z_M), M) = \mathbf{L}(U_{(1)}, \ldots, U_{(M')}, M') \tag{2.4}$$

where $M' \equiv \#\{i : U_i \le \pi\}$. Furthermore, by definition of $p\,(j, 1-\alpha)$, $p\,(n, 1-\alpha) = 1$ and

$$P\,[U_{(j+1)} \le p\,(j, 1-\alpha)] = 1-\alpha, \tag{2.5}$$

$$\text{for } 0 \le j \le n-1\,.$$

Now suppose $\pi = 1$ so that $M = n$. Then, we see that $\bar{\mu}_{ST}$ simply replaces each $1 - G\,(z_{j+1})$ by its $1-\alpha$ upper prediction bound $p\,(j, 1-\alpha)$.

We now prove,

THEOREM 1. *i) If G is continuous*

$$P\,[\mu \le \bar{\mu}_{ST}] \ge (1-\alpha)^{n+1}\,. \tag{2.6}$$

If $\pi = 1$ or $n = 1$ we can replace $(1-\alpha)^{n+1}$ by $(1-\alpha)^n$.

ii) If G is a point mass or $\pi = 1$ and G concentrates on at most 2 points, $\bar{\mu}_{ST}$ is conservative.

Proof. By (2.2) and (2.3),

$$P\,[\mu \le \bar{\mu}_{ST}] = P\,[\pi \sum_{j=0}^{M-1} (z_j - z_{j+1})(1 - G\,(z_{j+1}) - p\,(j, 1-\alpha)) \tag{2.7}$$

$$\le \pi z_M \{(p\,(M, 1-\alpha) - 1) + \frac{(1-\pi)}{\pi z_M} \bar{\mu}_{ST}\}]\,.$$

Now, from (1.3)

$$\frac{(1-\pi)}{\pi} \frac{\bar{\mu}_{ST}}{z_M} \ge \left(\frac{1}{\pi} - 1\right) p\,(M, 1-\alpha)\,.$$

So $P\,[\mu \le \bar{\mu}_{ST}]$ is bounded below by

$$P\,[\sum_{j=0}^{M-1}(z_j - z_{j+1})(1 - G(z_{j+1}) - p\,(j,1-\alpha)) \le 0,\quad p\,(M,1-\alpha) \ge \pi\,]$$

and a fortiori by,

$$P\,[\max_{1\le j\le M}(1 - G\,(z_j) - p\,(j,1-\alpha)) \le 0,\quad M \ge k\,(\pi)\,] \tag{2.8}$$

for the appropriate integer $k\,(\pi)$. Applying (2.4) we see that (2.8) can be bounded below by,

$$P\,[\max_{1\le j\le n}(U_{(j)} - p\,(j,1-\alpha)) \le 0,\quad M' \ge k\,(\pi)\,]. \tag{2.9}$$

But from Marshall and Proschan (1966), since the variables

$$(1\,(U_1 \le p\,(j,1-\alpha)),\ 1 \le j \le n,\quad 1\,(U_1 \le \pi))$$

are positively dependent we can conclude that

$$P\,[\max_{1\le j\le n}(U_{(j)} - p\,(j,1-\alpha)) \le 0,\quad M' \ge k\,(\pi)\,] \tag{2.10}$$

$$\ge \{\prod_{j=1}^{n} P\,[U_{(j)} \le p\,(j,1-\alpha)]\}\,P\,[M' \ge k(\pi)\,] = (1-\alpha)^{n+1}.$$

If $\pi = 1$, $k\,(\pi) = n$, $P\,[M' \ge n\,] = 1$ and the second statement of (i) follows. The case $n = 1$, $\pi < 1$ can be calculated directly. For (ii) note that if G is a point mass at z,

$$P\,[\mu \le \bar{\mu}_{ST}\,] = P\,[p\,(0,1-\alpha) + z\,(p\,(M,1-\alpha) - p\,(0,1-\alpha)) \ge \pi z\,]$$

$$\ge P\,[p\,(M,1-\alpha) \ge \pi\,] = 1-\alpha.$$

The case $\pi = 1$, G concentrates on two points is argued similarly. The theorem follows. □

This bound is obviously grossly inadequate since a great deal is lost in the passage from (2.7) to (2.8) and (2.8) to (2.10). That the situation is actually much better is suggested by what happens if π and G are kept fixed and $n \to \infty$.

THEOREM 2. *For all P,*

$$\bar{\mu}_{ST} = \bar{T} + \frac{c\,(P)}{n^{1/2}}\,z_{1-\alpha} + o_p\,(n^{-1/2}) \tag{2.11}$$

where, $\Phi(z_{1-\alpha}) = 1 - \alpha$ *and* Φ *is the standard normal d.f., and*

$$c(P) = \int_0^1 G^{-1}(1-t)\frac{\pi(1-2\pi t)}{2(\pi t(1-\pi t))^{1/2}}\,dt\,. \tag{2.12}$$

Further,

$$c^2(P) \geq \pi\left(\int_0^1 z^2\,dG - \pi\left(\int_0^1 z\,dG(z)\right)^2\right) \tag{2.13}$$

with = iff G concentrates on at most 2 points.

Note that (2.12) and (2.13) imply that,

$$\lim_n P[\mu \leq \bar{\mu}_{ST}] \geq 1 - \alpha$$

with strict inequality unless the distribution of T_1 concentrates on at most 3 points one of which is 0. Recall that by theorem 1 (ii), $P[\bar{\mu}_{ST} \geq \mu] \geq 1 - \alpha$ holds for all n if the distribution of T_1 concentrates on at most 2 points. The key argument in the proof of (2.13) is due to Y. Ritov.

Proof. Note that,

$$P[U_{(j+1)} \geq c] = P\left[\sum_{i=1}^{j+1} E_i \geq c\sum_{i=1}^{n+1} E_i\right] \tag{2.14}$$

$$= P\left[\sum_{i=1}^{j+1}(1-c)(E_i - 1) - \sum_{i=j+1}^{n+1} c(E_i - 1) \geq c(n+2) - (j+1)\right]$$

where $E_1, \ldots, E_n$ are independent standard exponential. Suppose $j \leq n(1-\varepsilon)$, $\varepsilon > 0$. Write $c = c_j = j{+}1/n{+}2\,(1 + v/(j+1)^{1/2})$ where $v = O(1)$). Then,

$$\sigma_j^2 = (1-c)^2(j+1) + c^2(n-j+1) = \Omega(j+1)$$

where Ω denotes order. Let,

$$K_{rj} = \sigma_j^{-r}\{(1-c)^r(j+1) + (-c)^r(n-j+1)\}K_r$$

where K_r is the rth cumulant of E_1. Then,

$$K_{3j} \equiv 2\sigma_j^{-3}\{-c^3(n+2) + (j+1)(3c^2 - 3c + 1)\}$$

$$= \Omega(j^{-1/2})$$

$$K_{4j} = 9\sigma_j^{-4}\{e^4(n+2) + 2(j+1)(3c^2 - 2(c + c^3))\}$$

$$= \Omega(j^{-1}).$$

and so on. Note that

$$\frac{\sigma_j^2}{j+1} = \frac{(n+2)}{(j+1)} c^2 - 2c + 1 \tag{2.15}$$

$$= -\frac{(j+1)}{n+2} + \frac{v^2}{n+2} + 1 \equiv A_2(\frac{j+1}{n+2}, v, (n+2)^{-1/2}).$$

We can similarly write,

$$j^{\frac{r-2}{2}} K_{rj} = A_r(\frac{j+1}{n+2}, v, (n+2)^{-1/2}) \tag{2.16}$$

where $A_r(\cdot,\cdot,\cdot)$ is entire in its arguments and $A_r(u,v,0) \equiv A_r(u)$ doesn't depend on v. By standard results on Edgeworth expansions, see for example Bhattacharya and Ranga Rao (1975),

$$P[U_{(j+1)} \geq c] = 1 - \Phi(v A_2^{-1/2}) \tag{2.17}$$

$$- \phi(v A_2^{-1/2}) \sum_{i=1}^{k} (j+1)^{-1/2} B_i(v, \frac{j+1}{n+2}, (n+2)^{-1/2})$$

$$+ O(j^{-\frac{(k+1)}{2}})$$

uniformly in $|v| \leq M$, $n^\delta \leq j \leq n(1-\varepsilon)$, $\delta < 1/4$ where $B_i(\cdot,\cdot,\cdot)$ are entire. Take $k > 3/\delta - 1$ so that the remainder in (2.17) is $o(n^{-3/2})$ for $j \geq n^\delta$. Let v, j range freely subject to $n^\delta \leq j \leq n/2$, $|v| \leq M$. By (2.17) we deduce, if $p_j \equiv \frac{j+1}{n+2}$;

$$p(j, 1-\alpha) = p_j + \frac{z_{1-\alpha}}{n^{1/2}} [p_j(1-p_j)]^{1/2} + O(j^{-1}).$$

Writing $v = z_{1-\alpha}[p_j(1-p_j)]^{1/2} n^{-1/2} + w(j+1)^{-1}$ and continuing in this fashion we can deduce that if $r_j \equiv (p_j)^{1/2}$

$$p(j, 1-\alpha) = p_j + \frac{z_{1-\alpha}}{n^{1/2}} C_0(r_j) + \frac{C_1(r_j)}{n} + \frac{C_2(r_j)}{n^{3/2}} + o(n^{-3/2}) \tag{2.18}$$

where $C_0(r) = r(1-r^2)^{1/2}$, and C_1 and C_2 are smooth. Therefore,

$$p(j,1-\alpha) - (p(j-1,1-\alpha) = \frac{1}{n+2} + \frac{z_{1-\alpha}}{n^{1/2}} \frac{C_0'(r_j)}{2} [r_j(1-r_j^2)]^{-1/2}(1+o(1))$$

$$+ \; o\,(n^{-3/2}).$$

Suppose $\pi < 1$. If F_n is the empirical of $U_1, \ldots, U_n$,

$$P\,[\,U_{(j+1)} \geq c_n \frac{(j+1)}{n}\,] \;=\; P\,[\,\frac{F_n\,(U_{(j+1)})}{U_{(j+1)}} \leq \frac{1}{c_n}\,]$$

$$\leq \; P\,[\,\inf\{\frac{F_n\,(x)}{x} : x \geq U_{(1)}\} \leq \frac{1}{c_n}\,]$$

$$\to \; 0 \;\text{ if }\; c_n \to \infty \;\text{ by }\; (6),\,(7)\; p.\,345$$

of Shorack and Wellner (1986). We hence obtain,

$$p\,(j, 1-\alpha) \;=\; O\,(\frac{j}{n}) \tag{2.19}$$

uniformly in j.

Then

$$\hat{\mu}_{ST} \;=\; \sum_{j=n^{\delta}}^{m} z_j\,[\,p\,(j, 1-\alpha) - p\,(j-1, 1-\alpha)\,] \tag{2.20}$$

$$+ \; O\,(n^{2\delta-1}).$$

By (2.19), if $\hat{\pi} \equiv \frac{m}{n}$,

$$\hat{\mu}_{ST} \;=\; \frac{1}{n}\sum_{j=1}^{m} z_j\,(1 + \frac{z_{1-\alpha}}{n^{1/2}}\,\frac{(1-2p_j)}{2}\,[\,p_j\,(1-p_j)\,]^{-1/2}\,(1 + o\,(1))) \tag{2.21}$$

$$+ \; o_p\,(n^{-1/2})$$

$$= \; \bar{T} + \frac{z_{1-\alpha}}{n^{1/2}} \int_0 G_n^{-1}(1-t)\,\frac{(1-2\hat{\pi}t)}{2}\,[\,\hat{\pi}t\,(1-\hat{\pi}t)\,]^{-1/2}\,(1 + o\,(1))\,dt \; + \; o_p\,(n^{-1/2})$$

where G_n is the empirical df of $V_1, \ldots, V_n$.

Since, with probability 1, $G_n^{-1}(t)$ converges uniformly to $G^{-1}(t)$ and $\hat{\pi} \to \pi$ we can apply dominated convergence to obtain (2.11) for $\pi < 1$. If $\pi = 1$ we carry through a similar argument for the upper tail $j \geq (1-\varepsilon)\,n$, upon noting that

$$P\,[\,U_{(j+1)} \geq c\,] \;=\; P\,[\,1 - U_{(n-j)} \geq c\,] = 1 - P\,[\,U_{(n-j)} \geq 1 - c\,]$$

or

$$p(j, 1-\alpha) = 1 - p(n-j-1, \alpha).$$

Finally we give Y. Ritov's argument for (2.13). Let

$$a(s) \equiv -G^{-1}(1-s).$$

Then, by integration by parts and Fubini,

$$c^2(P) = \int_0^1\int_0^1 [\pi u(1-\pi u)\pi v(1-\pi v)]^{1/2} da(u)\, da(v).$$

Similarly,

$$\sigma^2(P) = \pi\int_0^1 [G^{-1}(1-s)]^2 ds - \pi^2 | \int_0^1 [G^{-1}(1-s)]\, ds\,]^2 \tag{2.22}$$

$$= \pi\int_0^1\int_s^1 da(u)\int_s^1 da(v)\, ds$$

$$- \pi\int_0^1\int_s^1 da(u)\int_t^1 da(v)\, ds\, dt$$

$$= \pi\int_0^1\int_0^1 [(\pi u \vee \pi v) - \pi u\, \pi v\,]\, da(u)\, da(v)$$

where $\vee$ denotes max. But, if $u \leq v$,

$$[u(1-u)v(1-v)]^{1/2} \geq u(1-v) \tag{2.23}$$

with equality iff $u = v$. Comparing (2.21) and (2.22) we see that (2.13) follows and further that equality holds iff G^{-1} takes on at most two values or equivalently G concentrates on at most two points. The theorem follows. □

3 Some alternatives to the Stringer bound

It is not difficult to obtain bounds which can be *proved* to be conservative under the Stringer assumptions. Unfortunately these bounds tend to be even more conservative in practice than the Stringer bound. Here are two examples.

3.1. The Hoeffding bounds

Bickel, Godfrey and Neter (1989) discuss the following procedure. Hoeffding (1962) shows that if $0 \le T \le 1$, $\mu = E(T)$ then $P[\bar{T} > a] \le V(a,\mu)$, $\mu \le a < 1$ where $V(a,\mu) = ((1-\mu)/(1-a))^{n(1-a)}(\mu/a)^{na}$. $V(a,\mu)$ is just

$$\max\{\inf\{e^{-bt} E_P e^{bT} : b \ge 0\} : P \text{ concentrating on } [0,1], E_P T = \mu\}$$

which is achieved for $P[T=0] = 1 - P[T=1] = 1-\mu$. Note that $V \downarrow$ in a for fixed μ, $\uparrow$ in μ for fixed a. Now define $a(\mu)$ by,

$$V(a(u),\mu) = \alpha, \quad 0 < \mu \le \alpha^{1/n}$$

$$a(\mu) = 1, \quad \alpha^{1/n} < \mu \le 1.$$

Let

$$\bar{\mu} \equiv 1 - a^{-1}(1-\bar{T}). \tag{3.1}$$

Equivalently,

$$V(\bar{T};\bar{\mu}_H) = \alpha \tag{3.2}$$

since $V(1-a, 1-\mu) = V(a,\mu)$.

Then, by a standard argument,

$$P[\mu \le \bar{\mu}_H] \ge 1-\alpha. \tag{3.3}$$

It is in retrospect not surprising that this bound though used successfully for probabilistic purposes is extremely conservative. For n large, $\sigma^2 = \mathrm{Var}(T_1) > 0$,

$$P[n^{1/2}\frac{(\bar{T}-\mu)}{\sigma} > z] \to 1-\Phi(z) \sim \frac{1}{z(2\pi)^{1/2}} e^{-\frac{z^2}{2}}$$

for z large. On the other hand, $V(\mu + \frac{z}{n^{1/2}}, \mu) \to e^{2z^2}$. That is, V is conservative because it replaces σ^2 by the worst case 1/4 and the normal tail $1-\Phi(z)$ by $e^{-z^2/2}$.

3.2. The Kolmogorov Smirnov bound

R. Pyke has suggested the bound,

$$\bar{\mu}_{RP} = \bar{T} + d_\alpha^+$$

where,

$$P\,[\sup_x (F_n(x) - F(x)) \;\le\; d_\alpha^+] \;\ge\; 1 - \alpha$$

and F_n is the empirical d.f. of $T_1, \ldots, T_n$. Then,

$$P\,[\mu \le \bar{\mu}_{RP}] \;=\; P\,[\int_0^1 (F_n(x) - F(x))\,dx \;\le d_\alpha^+]$$

$$\ge\; P\,[\sup_x (F_n(x) - F(x)) \;\le\; d_\alpha^+]\,.$$

This bound shares the asymptotic extreme conservativeness of the Springer and Hoeffding bounds since

$$n^{1/2} d_\alpha^+ \;\to\; (\frac{1}{2}\log\frac{1}{\alpha})^{1/2},$$

so that this bound replaces $1 - \Phi(z)$ by $\exp(-2z^2)$ Alternatively we may seek approximate bounds which will be tighter and yet reasonably conservative. There are a number of Bayesian and other parametric proposals available — see Cox and Snell (1979) for example. But none of these seems to behave reliably if the distribution of T does not belong to the model.

The normal approximation bound, $\bar{\mu}_N \equiv \bar{T} + z_{1-\alpha} s n^{-1/2}$ where $s^2(n-1)/n$ is the sample standard deviation is, of course, asymptotically correct but is known to behave poorly (and is undefined for $m = 0$). A number of bootstrap alternatives are available — see diCiccio and Romano (1988). As an example we consider the ''nonparametric tilting'' bound which they show is ''second order asymptotically correct''.

Given $T_1, \ldots, T_n$, let $\{P_S^*\}$ be the exponential family of distributions placing mass proportional to e^{ts} on $t = T_1, \ldots, T_n$. Let $T_1^*, \ldots, T_n^*$ be a sample of size n from P_S^*. Let

$$u(s) \;=\; \int t\, dP_S^*(t) \;=\; \frac{\Sigma_{i=1}^n T_i\, e^{sT_i}}{\Sigma_{i=1}^n e^{sT_i}}\,.$$

Define $\hat{s}$ by,

$$P_{\hat{s}}^*[\bar{T}^* \ge \bar{T}] \;=\; 1 - \alpha$$

and

$$\bar{\mu}_{tilt} = \mu(\hat{s}).$$

That is, $\bar{\mu}_{tilt}$ is the $1-\alpha$ UCB for μ when sampling from the exponential family $\{P_s^*\}$. Calculation of $\bar{\mu}_{tilt}$ requires simulation of $\bar{T}^*$ under P_s for a range of values of s.

A natural simplification is to replace $P_s^*[\bar{T}^* \geq a]$ by its Hoeffding lower bound $B(a,s)$ and then solve, if possible, $B(\bar{T},s) = 1-\alpha$ to get $\hat{\hat{s}}$ and let $\bar{\mu} = \mu(\hat{\hat{s}})$. Compromises between the bootstrap and Hoeffding bounds such as this one are under investigation.

The essential difficulty of this problem is that M is typically moderate even though n is large so that M is approximately Poisson rather than normal and we are not close to asymptopia. It is this set of circumstances that the Stringer bound seeks to capture.

The following bound is proposed as a compromise between the Stringer and Gaussian bounds behaving like Stringer for M small and like the Gaussian bound for M large. Our point of departure is to write,

$$\sum_{i=1}^{n} T_i = \sum_{i=1}^{M} V_i ,$$

where M has a binomial (n,π) distribution and V_i has the conditional distribution G of T_i given $T_i > 0$.

1) We estimate π not by the bootstrap which would be M/n but the larger $p(M,1-\alpha)$, where $p(M;1-\alpha)$ is as defined previously.

2) We estimate the distribution G by $\hat{G}$, the empirical distribution of the positive T_i if $M > 0$. If $M = 0$ it is conservative to take $\hat{G}$ point mass at 1.

For any t we accordingly estimate $P[\sum_{i=1}^{n} T_i \geq n\mu - nt]$ by, $P^*[\sum_{j=1}^{M^*} V_i^* \geq np(M,1-\alpha)\bar{V} - nt]$ where, under P^*, M^* has a binomial $(n,p(M,1-\alpha))$ distribution and V_i^* are independent identically distributed $\hat{G}$. We then can solve, if $M = m$, for $\hat{t}_\alpha$,

$$P^*[\sum_{i=1}^{M^*} V_i^* \geq np(m,1-\alpha)\bar{V} - nt] \geq 1-\alpha \tag{3.1}$$

and $P^*[\sum_{i=1}^{M^*} V_i^* > np\,(m, 1-\alpha)\bar{V} - nt\,] < 1-\alpha.$

We then use

$$\bar{\mu} = \bar{T} + \hat{t}_\alpha.$$

as our bound.

1) If $m = 0$, $V^* \equiv 1$, $A \equiv [M^* \ge np\,(0, 1-\alpha) - nt\,]$

$$\begin{aligned} P^*[A\,] &= (1-(1-p\,(0,1-\alpha))^n)P^*[A\,|m^* \ge 1] \\ &= (1-\alpha)P^*[A\,|m^* \ge 1] \end{aligned}$$

since $p\,(0, 1-\alpha) = 1-\alpha^{1/m}$. The second term is $= 1$ iff $nt \ge np\,(0, 1-\alpha)$, that is, iff $\bar{\mu} = \hat{t}_\alpha = p\,(0, 1-\alpha)$

2) If $m = 1$,

$$\hat{t}_\alpha = V_1(\frac{1}{n} + p\,(1, 1-\alpha) - \frac{\hat{k}_\alpha}{n}) \tag{3.2}$$

where $P^*[M^* \ge \hat{k}_\alpha] \ge 1-\alpha$, if $M^* \sim bin\,(n, p\,(1, 1-\alpha))$

$$P^*[M^* > \hat{k}_\alpha] < 1-\alpha$$

3) For $m \ge 2$ we can approximate further to obtain a bound in closed form. Note that,

$$E^*(\sum_{i=1}^{M^*} V_i^*) = np\,(m, 1-\alpha)\,\bar{V}$$

$$\mathrm{Var}^*(\sum_{i=1}^{M^*} V_i^*) \le n\,\{p\,(m, 1-\alpha)\,s_v^2 + p\,(m, 1-\alpha)\,(1-p\,(m, 1-\alpha))\,\bar{V}^2\}$$

where,

$$s_v^2 \equiv \frac{1}{m-1}\sum_{i=1}^{m}(V_i - \bar{V})^2.$$

This leads to

$$\bar{\mu} = \bar{T} + \frac{z_{1-\alpha}}{n^{1/2}}(p\,(m, 1-\alpha)\,[s_v^2 + (1-p\,(m, 1-\alpha))\,\bar{V}^2])^{1/2}, \tag{3.3}$$

by (2.18) As $n \to \infty$, $p\,(M, 1-\alpha) \xrightarrow{P} \pi$, $s_v^2 \to \mathrm{Var}(V)$, $\bar{V} \to E\,(V)$ and the bound is asymptotically correct. Since $M/n \le p(M, 1-\alpha)$ we expect the bound to be conservative.

P. Lorentziadis has carried out a small simulation of this approximate bound with encouraging results. In all cases $n = 100$, $\alpha = .05$ and 1000 simulations were performed. We took $\pi = .06, .12$. The distributions G considered were:

1) Uniform $(0, 1)$

2) $G = \rho\{1\} + (1-\rho)\{t\}$ a mixture of point mass at 1 with point mass at t with probabilities ρ and $1-\rho$. We used,

   a) $\rho = .5$, $t = .5$

   b) $\rho = .9$, $t = .1$.

For comparison we table,

c_{sT} - the coverage probability of the Stringer bound

c_N - the coverage probability of our bound

$m_{st}, (m_N)$ - the average overshoot of the Stringer (new) bounds when they cover. That, is $m_{ST} \cong E(\bar{\mu}_{ST} - \mu)_+$ and m_N is defined similarly

		c_{ST}	m_{ST}	c_N	m_N
Situation	1	1.00	.05	.95	.03
	2a)	1.00	.06	.96	.04
	b)	1.00	.07	.98	.05

$\pi = .06$

		c_{ST}	m_{ST}	c_N	m_N
Situation	1	1.00	.07	.95	.04
	2a)	.99	.08	.96	.06
	b)	.99	.09	.97	.07

$\pi = .12$

Further investigation of this bound and the "exact" form (without Gaussian approximation for $m \geq 2$) is envisaged.

Acknowledgement: I thank P. Lorentziadis for the simulations and J. Neter for helpful conversations.

References

Bhattarcharya, R. and Ranga, Rao R. (1976). Normal Approximations and Asymptotic Expansions. J. Wiley, New York.

Bickel, P.J., Godrey, J., Neter J. & Clayton, H. (1989). Hoeffding bounds for monetary unit sampling in auditing. Contributed paper I.S.I. *Meeting*, Paris.

Cox, D.R. and Snell, E. (1979). On sampling and the estimation of rare errors. *Biometrika* **66**, 124-132.

diCiccio, T. & Romano, J. (1988). A review of bootstrap confidence intervals: with discussion. *JRSSB* **50**, 338-370.

Hoeffding, W. (1963). Probability inequalities for sums of random variables. *JASA* **58**, 13-29.

Shorack, G. & Wellner, J. (1986). Empirical Processes with Applications to Statistics. J. Wiley, New York.

Statistical Models and Analyses in Auditing. *Statistical Science* **4**, 2-33.