Biometrika (2000), **87**, 3, *pp.* 689–694

Miscellanea

Inference on rare errors using asymptotic expansions and bootstrap calibration

BY ROELOF HELMERS
CWI, PO Box 94079, 1090 GB Amsterdam, The Netherlands
helmers@cwi.nl

SUMMARY

The number of items in error in an audit population is usually quite small, whereas the error distribution is typically highly skewed to the right. For applications in statistical auditing, where line-item sampling is appropriate, a new upper confidence limit for the total error amount in an audit population is obtained. Our method involves an empirical Cornish–Fisher expansion in the first stage; in the second stage we employ the bootstrap to calibrate the coverage probability of the resulting interval estimate.

Some key words: Auditing; Cornish–Fisher expansion; Finite population; Line item sampling; Nonstandard mixture; Poisson distribution; Rare error; Superpopulation model.

1. INTRODUCTION

The problem investigated in this paper arises in statistical auditing. Consider a finite population of N items with recorded values $y_1, \ldots, y_N$, the 'book amounts'. Suppose that the items may be subject to unknown errors $e_1, \ldots, e_N$ and that $x_1, \ldots, x_N$ are the 'true' values, the 'audited' amounts of N items. Thus, $e_i = y_i - x_i$ denotes the 'error amount' corresponding to the book value y_i of the ith item or audit unit, $i = 1, \ldots, N$. However, it is known a priori that most of the e_i's are zero, and the auditor's problem is to give a $(1-\alpha)$ upper confidence bound for the total population error

$$D = \sum_{i=1}^{N} e_i \tag{1·1}$$

when a random sample S of book amounts of size n drawn without replacement from the population $\{y_1, \ldots, y_N\}$ is available, and e_i, for $i \in S$, denote the errors observed by the auditor in the recorded values Y_i, for $i \in S$. Clearly $\sum_{i \in S} e_i$ is the total error amount in the sample, and $\hat{D}_n = Nn^{-1} \sum_{i \in S} e_i$ is an unbiased estimator of D, our parameter of interest.

Let p denote the small probability that e_i is nonzero, and let M be the number of items in the sample $Y_1, \ldots, Y_n$ with error. Clearly

$$\hat{D}_n = \frac{N}{n} \sum_{j=1}^{M} V_j, \tag{1·2}$$

where the V_j's are the observed nonzero error amounts in the sample. In typical applications errors are rare, that is p is close to zero, and the sample size n is small compared with the size of the population N. In such cases one may impose a superpopulation model on $(y_1, \ldots, y_N)$, where $y_i = x_i + e_i$, for $i = 1, \ldots, N$, by assuming that the e_i's are independent random variables with a common

distribution of nonstandard mixture type (Tamura, 1989)

$$F = pG + (1-p)\delta_0, \tag{1·3}$$

where δ_0 denotes the degenerate distribution which puts all its probability mass at zero. Clearly the V_j's constitute a random sample of random size M from the unknown nonzero error amount distribution G, while M is the random number of nonzero errors present in a random sample of size n from F. In such cases one may assume that M is Po(v)-distributed, with unknown parameter $v = np$; in addition M is assumed to be independent of the V_j's. The Poisson approximation for M works well, provided the error rate p is small. There is no use for the classical requirement that $v = np$ is fixed, while $p \downarrow 0$ and $n \to \infty$. On the contrary, in the present paper we let $v = v_n$ approach infinity, as $n \to \infty$, whereas p is assumed to be small but fixed. We refer to Barbour, Holst & Janson (1992) for an excellent account of the theory of Poisson approximations.

Let $\mu = \int x\, dG(x)$ denote the nonzero mean of G; G is just the distribution of V_1, that is the conditional distribution of an error amount e, given that $e \neq 0$. Since $n \ll N$ one may argue that for our purposes D, the total population error in the finite population under consideration, can be replaced by $E_F D = Nn^{-1}v\mu$, under random sampling from the superpopulation error distribution F; the only exception would be the case that v is extremely small, but we rule out this case here. The problem is now that of finding a $(1-\alpha)$ upper confidence limit for $E_F D$. Note that G is typically highly skewed to the right.

In statistical auditing items are often selected without replacement with probability proportional to recorded book values, e.g. by applying dollar-unit sampling. In the present paper, however, we employ simple random sampling without replacement, i.e. audit-unit or line-item sampling. This appears more convenient in a variety of situations where it is equally important to ascertain the correct value for each audit unit, and each y_i should have equal chance of being included in our sample. For instance, in social security, payments of disability or unemployment benefits should be correct, irrespective of whether the benefit is a large or a small amount. Also in tax examinations and other audit applications in the public sector the auditor employs line-item or audit-unit sampling (Tamura, 1989, p. 6).

In this paper we establish an upper confidence limit with confidence level at least equal to $(1-\alpha)$ for the total population error D using asymptotic expansions and bootstrap calibration. Our focus is on the important situation that errors are rare and the nonzero error distribution is highly skewed. In realistic cases G may consist of a finite mixture of light-tailed distributions, such as the exponential. It is well known that such mixtures are hard to distinguish from heavy tail models (Jensen, 1995, Ch. 7). Hence one should not only correct for skewness but for kurtosis as well. Our method will give a much better one-sided confidence interval for D than the traditional normal approximation. However, no method for setting confidence limits for D will work in all cases. For example, the method proposed in this paper would not be suitable for cases with $M = 0$ or 1 and sample sizes as small as $n = 100$, say.

2. ASYMPTOTIC EXPANSIONS

As the normal approximation typically behaves poorly in audit populations one might try first to improve upon this by employing Cornish–Fisher expansions, adapting for skewness and kurtosis by estimating the third and fourth cumulants appearing in the Cornish–Fisher expansion from the observed nonzero error amounts at hand. However, one cannot really expect the empirical Cornish–Fisher expansion to work well in most instances, as our estimates of the third and fourth cumulants are by necessity highly variable, because the number of nonzero errors is usually quite small. We therefore employ bootstrap calibration to extend the range of validity of our method; see § 3.

We will assume throughout that p is fixed but close to zero, so that the Poisson approximation for M is applicable, while at the same time the sample size n approaches infinity. Thus the expected number of nonzero errors in the sample $E(M) = v = np$ also gets large in the asymptotics. A simple

calculation gives

$$E_F \hat{D}_n = \nu\mu \frac{N}{n}, \quad \sigma_F^2(\hat{D}_n) = \nu(\mu_2 - p\mu^2)\frac{N^2}{n^2} \sim \nu\mu_2 \frac{N^2}{n^2}, \tag{2·1}$$

provided G is nondegenerate, that is $\mu_2 > 0$, where $\mu = E_G V_1$, $\mu_2 = E_G V_1^2$; the relative error commited in the approximation (2·1) is of order p, as $p \downarrow 0$. The third and fourth cumulants κ_{3n} and κ_{4n} of $\hat{D}_n$ are

$$\kappa_{3n} = E_F(\hat{D}_n - E_F\hat{D}_n)^3/\sigma_F^3(\hat{D}_n) \sim \kappa_3 = \frac{\tilde{\mu}_3 + 3\sigma^2\mu + \mu^3}{\nu^{1/2}\mu_2^{3/2}} \tag{2·2}$$

with $\sigma^2 = \sigma_G^2(V_1)$, $\tilde{\mu}_3 = E_G(V_1 - \mu)^3$ and

$$\kappa_{4n} = E_F(\hat{D}_n - E_F\hat{D}_n)^4/\sigma_F^4(\hat{D}_n) - 3 \sim \kappa_4 = \frac{\tilde{\mu}_4 + 4\mu\tilde{\mu}_3 + 6\sigma^2\mu^2 + \mu^4}{\nu\mu_2^2}, \tag{2·3}$$

where $\tilde{\mu}_4 = E_G(V_1 - \mu)^4$. The errors committed in the approximations in (2·2) and (2·3) are of orders $(p/n)^{1/2}$ and $1/n$ respectively, and relative error of order p. The quantities κ_3 and κ_4 are easily checked to be exactly equal to the third and fourth cumulants of $\sum_{j=1}^M V_j$, where the V_j's denote a random sample from G, with Po(ν)-distributed random sample size M.

Define studentised statistics $S_{1,n}$ and $S_{2,n}$ by

$$S_{1,n} = \frac{\hat{D}_n - \nu\mu N/n}{(\sum_{j=1}^M V_j^2)^{1/2} N/n}, \quad S_{2,n} = \frac{\hat{D}_n - \nu\mu N/n}{Nn^{-1/2}\hat{s}}, \tag{2·4}$$

where $\hat{s}^2 = n^{-1}\sum_{j=1}^n (e_j - \bar{e})^2$, with $\bar{e} = n^{-1}\sum_{j=1}^n e_j$. Note that

$$\mathrm{pr}_F(S_{1,n} \leqslant x) = \mathrm{pr}_F(S_{2,n} Q^{\frac{1}{2}} \leqslant x), \tag{2·5}$$

where

$$Q = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{\sum_{j=1}^M V_j^2}. \tag{2·6}$$

A simple argument yields

$$Q = 1 - \frac{M^2\bar{V}^2}{n\sum_{j=1}^M V_j^2}, \tag{2·7}$$

and it is easily verified that

$$\mathrm{pr}_F(S_{1,n} \leqslant x) \geqslant \mathrm{pr}_F(S_{2,n} \leqslant x) \tag{2·8}$$

for $x \geqslant 0$, while the reverse inequality holds for $x < 0$.

The distribution of $S_{2,n}$ is the distribution of the classical Student t-statistic $n^{\frac{1}{2}}(\bar{e} - p\mu)/\hat{s}$, based on a sample of size n from F. Of course $\int x\, dF(x) = p\mu$. Let $c_{2,n,\alpha}$ denote the $(1-\alpha)$th critical point of $S_{2,n}$, that is $\mathrm{pr}(S_{2,n} \leqslant c_{2,n,\alpha}) = 1 - \alpha$. In Example 1 of Hall (1988), one can find a Cornish–Fisher expansion of $c_{2,n,\alpha}$:

$$c_{2,n,\alpha} \sim u_\alpha + \left(\frac{2u_\alpha^2 + 1}{6}\right)\kappa_{3n} + u_\alpha\left\{-\frac{1}{12}\kappa_{4n}(u_\alpha^2 - 3) + \frac{5}{72}\kappa_{3n}^2(4u_\alpha^2 - 1) + \frac{1}{4}n^{-1}(u_\alpha^2 + 3)\right\}, \tag{2·9}$$

where $u_\alpha = \Phi^{-1}(1-\alpha)$, and κ_{3n} and κ_{4n} are the third and fourth cumulants of $\bar{e} = n^{-1}\sum_{i=1}^n e_i$, that is $\hat{D}_n$, under random sampling from F; $\sim$ refers here to the fact that we have deleted terms of smaller order than n^{-1}. The expansion (2·9) has a remainder of order $o(n^{-1})$ as n gets large, provided F possesses an absolutely continuous component and a fourth moment of F exists (Hall, 1987). Note that p is close to zero, but assumed to be fixed; that is, F is also fixed in the asymptotics, as required by Hall (1987, 1988), otherwise the nonsingularity requirement may cause problems.

Let now $c_{1,n,\alpha}$ denote the $(1-\alpha)$th critical point of $S_{1,n}$, that is $\mathrm{pr}(S_{1,n} \leqslant c_{1,n,\alpha}) = 1-\alpha$, where, compare with (2·4), $S_{1,n} = (\sum_{j=1}^{M} V_j - \nu\mu)/(\sum_{j=1}^{M} V_j^2)^{1/2}$. If $0 < \alpha < \frac{1}{2}$, then obviously $c_{1,n,\alpha} \leqslant c_{2,n,\alpha}$ for sufficiently large n. Hence, Hall's expansion (2·9) for $c_{2,n,\alpha}$ provides us with a conservative approximation for $c_{1,n,\alpha}$ for confidence levels $1-\alpha$ with $0 < \alpha < \frac{1}{2}$ for sufficiently large n, provided p is close to zero. It follows that an approximate upper confidence limit with confidence level at least equal to $(1-\alpha)$ for $E_F D = Nn^{-1}\nu\mu$ is

$$U = \hat{D}_n + c_{2,n,\alpha}\frac{N}{n}\left(\sum_{j=1}^{M} V_j^2\right)^{\frac{1}{2}}, \tag{2·10}$$

provided the error rate p is small enough and $0 < \alpha < \frac{1}{2}$. However, the upper bound (2·10) cannot be computed from the data, as the 'theoretical critical point' $c_{2,n,\alpha}$ (Hall, 1988) is unknown, because G is unknown. Hence we replace (2·10) by its empirical counterpart,

$$\hat{U} = \hat{D}_n + \hat{c}_{2,n,\alpha}\frac{N}{n}\left(\sum_{j=1}^{M} V_j^2\right)^{\frac{1}{2}}, \tag{2·11}$$

where $\hat{c}_{2,n,\alpha}$ is approximated by

$$\hat{c}_{2,n,\alpha} \sim u_\alpha + \left(\frac{2u_\alpha^2+1}{6}\right)\hat{\kappa}_3 + u_\alpha\left\{-\frac{1}{12}\hat{\kappa}_4(u_\alpha^2-3) + \frac{5}{72}\hat{\kappa}_3^2(4u_\alpha^2-1) + \frac{1}{4}M^{-1}(u_\alpha^2+3)\right\}, \tag{2·12}$$

with

$$\hat{\kappa}_3 = \frac{\tilde{\mu}_3^* + 3\tilde{\mu}_2^*\bar{V} + \bar{V}^3}{M^{-1}(\sum_{j=1}^{M} V_j^2)^{\frac{3}{2}}},$$

$$\hat{\kappa}_4 = \frac{\tilde{\mu}_4^* + 4\tilde{\mu}_3^*\bar{V} + 6\tilde{\mu}_2^*\bar{V}^2 + \bar{V}^4}{M^{-1}(\sum_{j=1}^{M} V_j^2)^2}, \tag{2·13}$$

where $\tilde{\mu}_l^* = M^{-1}\sum_{j=1}^{M}(V_j - \bar{V})^l$, for $l = 2, 3, 4$ and $\bar{V} = M^{-1}\sum_{j=1}^{M} V_j$. Clearly, the coverage probability of empirical Cornish–Fisher bound (2·11) satisfies the inequality

$$\lim_{n\to\infty} \mathrm{pr}_F(E_F D < \hat{U}) \geqslant 1-\alpha, \tag{2·14}$$

provided the error rate p is small enough and $0 < \alpha < \frac{1}{2}$. In (2·12) $\sim$ indicates that, in addition to the error terms already deleted in the previous steps, the random approximation error in (2·12), caused by replacing our Cornish–Fisher expansion by its empirical counterpart, is of smaller order in probability than ν^{-1}, as $\nu \to \infty$.

3. BOOTSTRAP CALIBRATION

The empirical Cornish–Fisher bound (2·11) is easy to compute. However, the coverage probability $\mathrm{pr}_F(E_F D < \hat{U})$, compare (2·14), may in fact not be at least equal to the nominal confidence level $1-\alpha$, as desired in finite samples. To remedy this defect one may employ bootstrap calibration (Beran, 1987; Hall & Martin, 1988). The idea is to estimate by means of resampling the coverage probability, with α replaced by λ, for a grid of values of λ in $(0, 1)$, and select the largest value $\hat{\lambda}$ for which the bootstrap estimate

$$P_n^*\left\{\hat{D}_n < \hat{D}_n^* + \hat{c}_{2,n,\lambda}^*\frac{N}{n}\left(\sum_{j=1}^{M^*} V_j^{*2}\right)^{\frac{1}{2}}\right\} \tag{3·1}$$

is at least $1-\alpha$. Here P_n^* refers to probability in our 'bootstrap world': conditionally given $(V_1, \ldots, V_M)$, a bootstrap resample $(V_1^*, \ldots, V_{M^*}^*)$ of size M^* is drawn with replacement from $(V_1, \ldots, V_M)$, where the random resample size M^* is a realisation of a Poisson distribution with parameter M; $\hat{c}_{2,n,\lambda}^*$ is $\hat{c}_{2,n,\alpha}$, see (2·12), with α replaced by λ and $(V_1, \ldots, V_M)$ by $(V_1^*, \ldots, V_{M^*}^*)$,

while $\hat{D}_n^* = Nn^{-1}\sum_{j=1}^{M^*} V_j^*$. We note in passing that $\hat{\lambda}$ may not exist in exceptional cases. However, in the simulations reported in § 4, $\hat{\lambda}$ could always be determined. The numerical grid of λ-values was taken to be equally spaced with respect to the corresponding u_λ-values with constant width 0·01. This amounts to differences between subsequent λ-values not bigger than 2×10^{-4} in our simulation. A minor difficulty arises when $M^* = 0$, that is there is no bootstrap sample and we simply delete such 'empty' resamples; accordingly the P_n^*-probability (3·1) is estimated in such cases by the number of times the inequality in (3·1) is valid divided by the number of 'nonempty' bootstrap samples. Note that, when $M \geqslant 5$, the probability that $M^* = 0$ is at most equal to $e^{-5} = 0{\cdot}0067$.

Obviously $\hat{\lambda}$ will typically be somewhat smaller than α, and the calibrated confidence bound

$$\hat{D}_n + \hat{c}^*_{2,n,\hat{\lambda}} \frac{N}{n} \left(\sum_{j=1}^{M} V_j^2 \right)^{\frac{1}{2}} \tag{3·2}$$

will usually be larger than (2·11), but the calibrated upper bound (3·2) possesses the beneficial property of having estimated confidence level at least equal to $1 - \alpha$. Our bootstrap estimate (3·1), with $\lambda = \alpha$, may be used as a diagnostic tool to check whether or not the empirical Cornish–Fisher bound already has the desired confidence level $\geqslant 1 - \alpha$, and calibration of the bound (2·11) would not be needed.

In contrast to (2·11), the bound (3·2) requires a lot of computation, as it involves extensive bootstrapping. In practice, however, bootstrap calibration will only be needed when the dataset at hand contains few errors and/or the observed nonzero error amounts in the sample contain one or more 'outliers'. Otherwise, it is to be expected that the computationally much simpler bound (2·11) will usually suffice. One may try to develop a practical guideline for the use of bootstrap calibration in our setting; see Young (1994, p. 411) for similar advice. In any case, the computationally very demanding double bootstrap technique is avoided as our starting interval (2·11) is a non-bootstrap interval. For this very reason we have not used the studentised bootstrap (Hall, 1988; Helmers, 1991) in the first stage, but instead relied on an empirical Cornish–Fisher expansion.

4. Simulations

In this section we briefly describe some Monte Carlo simulations for various audit populations of practical interest. The size of the finite population under consideration was set at $N = 5 \times 10^5$.

In our first simulation we take $p = 0{\cdot}02$, $n = 500$ and $G = \mathrm{Ex}(200)$, the exponential distribution with mean 200. The errors e_i are distributed according to the nonstandard mixture distribution

$$F = 0{\cdot}02\ \mathrm{Ex}(200) + 0{\cdot}98\delta_0. \tag{4·1}$$

This first example represents a relatively simple audit population. The parameter of interest D is replaced by $E_F D = Nn^{-1}\nu\mu$, which equals 2×10^6. The number of nonzero errors in our sample of size 500 from F is Poisson distributed with mean 10. The true coverage probability $\mathrm{pr}_F(E_F D < \hat{U})$, with $(1 - \alpha) = 0{\cdot}95$, was estimated accurately by Monte Carlo to be 0·938, using 5×10^5 samples from F. Next, on the basis of a single sample of size 500 from F, the bootstrap estimate (3·1), with $\lambda = \alpha$, of the coverage probability was computed, using $B = 5000$ bootstrap resamples, with random resample size $\mathrm{Po}(M)$, where M denotes the number of errors in the original sample from F. This procedure was repeated 2000 times. The average of these 2000 bootstrap estimates of the true coverage probability 0·938 equals 0·932, while a density plot of these estimates is given in Fig. 1(a). The graph shows that our bootstrap estimate for the coverage probability of the Cornish–Fisher bound reflects about 84% of the time the fact that our upper confidence limit (2·11) has a true confidence level somewhat smaller than 0·95, namely 0·938. Hence, computing (3·1), with $\lambda = \alpha$, yields a fairly reliable diagnostic for the validity of the Cornish–Fisher upper bound (2·11) in this case. Calibration is perhaps needed here, as 0·932 falls short of 0·95.

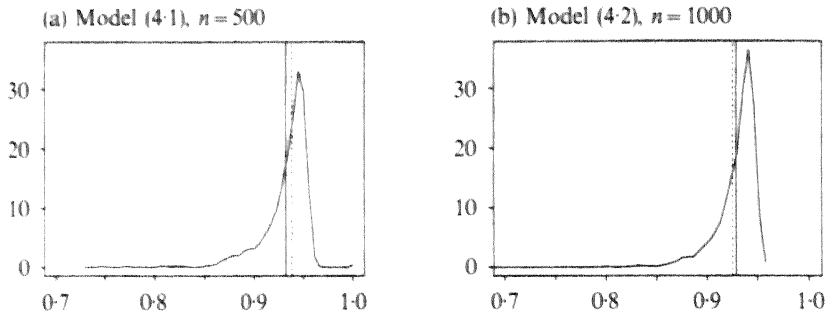

Fig. 1. Density of 2000 bootstrap estimates (3·1), with $\lambda = \alpha$, of true coverage probability. Dashed vertical line, true coverage probability; solid vertical line, average of 2000 bootstrap estimates.

In a second simulation we consider a more realistic nonstandard mixture:

$$F = 0{\cdot}02\ \mathrm{Ex}\left(\frac{100}{3}\right) + 0{\cdot}01\ \mathrm{Ex}\left(\frac{1000}{3}\right) + 0{\cdot}97\delta_0. \tag{4·2}$$

In this set-up we take into account the possibility of outliers among the observed nonzero error amounts, by assuming that G consists of a mixture of two exponentials, with means $\frac{100}{3}$ and $\frac{1000}{3}$ respectively. In the present example we take $n = 1000$ and $E_F D = Nn^{-1}(v_1\mu_1 + v_2\mu_2)$, with $v_1 = 20$, $\mu_1 = \frac{100}{3}$, $v_2 = 10$, $\mu_2 = \frac{1000}{3}$. Again $E_F D = 2 \times 10^6$. The number of nonzero errors in our sample of size 1000 from F is now Poisson distributed with mean 30; on average 10 of these will be outliers. The true coverage probability, with $(1 - \alpha) = 0{\cdot}95$, was estimated by Monte Carlo to be 0·925 using 5×10^5 samples from F. Next (3·1), with $\lambda = \alpha$, was estimated 2000 times, employing 2000 samples of size 1000 from (4·2) and using $B = 5000$ bootstrap resamples each time. The results are summarised in Fig. 1(b). The bootstrap diagnostic works well.

ACKNOWLEDGEMENT

This paper was written as part of a research project with the Statistical Audit Group of PricewaterhouseCoopers, Amsterdam. I thank Dick van der Hoeven for suggesting the problem. Comments of Harmen Ettema and John Haworth are gratefully acknowledged. Rob van der Horst from CWI carried out the simulations. I thank the editor and two referees for very valuable remarks.

REFERENCES

BARBOUR, A. D., HOLST, L. & JANSON, S. (1992). *Poisson Approximation*. Oxford Studies in Probability. Oxford: Clarendon.

BERAN, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74**, 457–68.

HALL, P. (1987). Edgeworth expansion for Student's t statistic under minimal moment conditions. *Ann. Prob.* **15**, 920–31.

HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals (with Discussion). *Ann. Statist.* **16**, 927–85.

HALL, P. & MARTIN, M. A. (1988). On bootstrap resampling and iteration. *Biometrika* **75**, 661–71.

HELMERS, R. (1991). On the Edgeworth expansion and the bootstrap approximation for a studentized U-statistic. *Ann. Statist.* **19**, 470–84.

JENSEN, J. L. (1995). *Saddlepoint Approximations*. Oxford: Clarendon.

TAMURA, H. (1989). Statistical models and analysis in auditing. Panel on Nonstandard Mixtures of Distributions. *Statist. Sci.* **4**, 2–33.

YOUNG, G. A. (1994). Bootstrap: more than a stab in the dark? (with Discussion). *Statist. Sci.* **9**, 382–415.

[*Received October* 1997. *Revised March* 1999]