Journal of Official Statistics, Vol. 20, No. 4, 2004, pp. 705–723

Confidence Intervals for Proportions Estimated from Complex Sample Designs

Alistair Gray[1], *Stephen Haslett*[2], *and Geoffrey Kuzmicich*[3]

Korn and Graubard (1998) have suggested a method for producing confidence intervals for proportions estimated from a sample based on a complex sample design where the proportions are either very small or very large, or the sample size is small. Their method uses the exact binomial confidence intervals but with the sample size modified by dividing by the estimated design effect for the proportion in question. Statistics New Zealand wanted to use this method for the 1999 Gaming Survey (which has a different design to that for which the Korn and Graubard method was developed) and carried out a bootstrap analysis to see whether the coverage properties of such intervals were similar to those from the Korn and Graubard method. The article presents the results of this analysis.

Key words: Complex sample design; proportions; exact binomial confidence interval; bootstrap confidence interval; exchangeable bootstrap.

1. Introduction

In 1999 Statistics New Zealand carried out a survey on gaming for the New Zealand Department of Interval Affairs (from now on referred to as the 1999 NZGS). The principal statistics from this survey were the proportions of problem and pathological gamblers. These proportions are small: at the national level estimates were 1.3% and 0.5%, respectively. We were interested in using the method proposed by Korn and Graubard (1998) for producing confidence intervals, but were concerned that the design for the 1999 NZGS was very different (small psu's and large variation in weights) from the type of survey Korn and Graubard had used in their simulation study.

Since we had neither previous census nor previous sample data, we knew little about the subpopulation of interest, namely problem and pathological gamblers in New Zealand, except for the information contained in the one sample. Because the characteristic of interest in the survey was relatively rare, we had insufficient information to adequately simulate this population.

Hence we decided to carry out a bootstrap analysis on the survey data. As a result of this analysis we developed the theory of the exchangeable bootstrap. This new method compare favourably with the Korn and Graubard approach.

[1] Statistics New Zealand, now at Statistics Research Associates Ltd., PO Box 12 649, Thorndon, Wellington, New Zealand. Email: alistair@statsresearch.co.nz
[2] Massey University, PO Box 11-222, Palmerston North, New Zealand. Email: s.j.haslett@massey.ac.nz
[3] Statistics New Zealand, now at Ministry of Justice, PO Box 180, Wellington, New Zealand.
Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect those of Statistics New Zealand.

This is not a theory article with a single practical example but rather an exposition of a theoretical development arising from a particular survey where certain technical difficulties arose from sampling a comparatively rare characteristic. We have consequently focused on sensitivity of the estimates to changes in the particular sample, rather than wider specification of artificial populations. Indeed, we take the view that the theory developed for the exchangeable bootstrap is more convincing than a simulation using artificial populations could be. In essence (as the theory we developed shows), if the simulation is set up with the appropriate exchangeable groups, the method will work very well, and if these groups are poorly chosen this will be obvious and the method will (as proven) give poor results. We see this as a useful diagnostic and not a failing, because even if the groupings are unknown, good choice of exchangeable groups will lead to "similarity" of sample mean and the average of the bootstrap means.

We believe our situation and solution best reflect actual requirements and constraints in government statistical agencies when implementing new surveys and forming estimates, as distinct from viewing the situation as an abstract research problem.

In Section 2 we discuss the New Zealand Gaming Survey sample design. In Section 3 we outline standard approaches for constructing confidence intervals for binomial data. Section 4 briefly summarizes bootstrapping for sample surveys. In Section 5 the theory of the exchangeable bootstrap is developed. We discuss producing bootstrap based confidence intervals for survey data in Section 6. Section 7 provides the results of our analysis and Section 8 comments in overview.

2. The 1999 Gaming Survey

The population for the 1999 NZGS was people aged 18 and over who were usually resident in New Zealand and who lived in private dwellings. Because the interviews for this survey were to be conducted over the telephone, for simplicity the frame for it was taken to be the telephone numbers listed in the Telecom New Zealand directory. Pilot testing suggested that a higher response rate could be achieved by sending a recruitment letter to the address belonging to the sampled telephone number a few days before the telephone interview was conducted.

The sample design was straightforward. The 18 Telecom New Zealand directories were used as strata. The stratum sizes varied from Auckland with a population of about 1,000,000 to the West Coast of the South Island with a population of about 23,000. The sample was allocated proportionally to the strata and selected by Telecom New Zealand using an algorithm to implement a simple random sample which was approved by Statistics New Zealand. The telephone number represented a cluster of people, typically one household. Of course a number of households had more than one telephone number associated with them: for example, for fax or internet traffic. Sampled households were asked for the number of telephone lines they had so that the probability of selection of the household could be correctly determined. One person per household (represented by the sampled telephone number) was selected using a modified Kish selection grid. In short, the design was a stratified single stage cluster design, where within each stratum a large number of psu's were selected and the psu's sizes were small. (The average number of people aged 18 and over living in private dwellings is about 2, and the maximum number is around 9.)

Sampling one unit in the ultimate cluster is mentioned by Kish (1965, p. 398) and is reasonably common in surveys carried out by official statistical agencies. The decision to sample one person per household was made when designing and implementing the 1999 NZGS, because there is greater benefit (from the point of view of accuracy for fixed cost) in sampling another household rather than sampling another individual within the same household. This does not allow unbiased estimation of variances, but only at a household level, and the bias effect for variance estimates taken overall is very small. Estimates of means or totals remain unbiased. In essence, the household in the 1999 NZGS may be considered as a cluster in the context of a proportionally allocated stratified sample of households. The variance within clusters is then necessarily an effect of considerably less importance than variability between households, as Cochran (1977) for example shows. What is much more important is knowing and using the number of eligible people per household for weighting purposes, as was done in the estimation phase of this survey.

The estimator chosen to produce estimates of such statistics as the proportions of problem and pathological gamblers was a calibration estimator using sex, age, and ethnic groups for the calibration variables. Straight poststratification to three-way table formed by these variables was not carried out because the population estimates of the cells of this three-way table were unreliable. Even if reliable estimates of the cells were available, it is likely such a fine breakdown would lead to more unstable estimates because the sample sizes in some of the cells were very small. Hence the choice of a calibration estimator.

Apart from the sample design differences between the Korn and Graubard study (explained below) and this study, a further difference is that we were using real data with biases rather than simulated populations. There is some bias in the frame. Firstly, although about 97% of New Zealand households have a telephone, this percentage drops to between 80% and 85% for lower socio-economic groups, in particular Māori and Pacific peoples. Secondly, at the time the survey was conducted about 10% of numbers were unlisted. Although some 9,300 telephone numbers were sampled only 6,452 interviews were completed. This response rate of about 75% is not as high as Statistics New Zealand usually achieves in its ongoing household surveys and may be attributable to the sensitive nature of the information sought or the interview mode. Furthermore, the response rate was not uniform across age groups or ethnic groups: for example, as is usual in household surveys in New Zealand, young males have higher nonresponse rates than average. Nevertheless, the 1999 NZGS achieved a response rate much larger than other gaming-related surveys carried out previously in New Zealand or elsewhere.

3. Confidence Intervals for Binomial Data

Statisticians required to produce confidence intervals for proportions from large surveys based on complex sample designs typically assume that the estimators are approximately normally distributed and so use normal confidence intervals. For example, suppose $\hat{p}$ is the estimated proportion and $\hat{v}_{\hat{p}}$ the estimated variance, then a $100(1-2\alpha)\%$ confidence interval would be $\hat{p} \pm z_\alpha \sqrt{\hat{v}_{\hat{p}}}$, where z_α is the $1 - \alpha$ quantile of the standard normal distribution.

If the sample size was very small taking into account the number of strata and primary sampling units (psu's) (so that the degrees of freedom were small), then one might use

instead of z_α in the above confidence interval the α critical value from a t-distribution with ν degrees of freedom, where ν is the number of psu's less the number of strata.

Of course, even in the simpler situation of simple random sampling, the symmetry of such confidence intervals is misleading with very small or very large proportions. So a common approach is to use a logit transformation of the proportion to produce the confidence interval (see Rust and Rao 1996). Specifically, as before, suppose $\hat{p}$ is the estimated proportion and $\hat{v}_{\hat{p}}$ the estimated variance. Carry out the following steps:

1. Calculate the logit, $\hat{L}$, of $\hat{p}$: $\hat{L} = \log(\hat{p}/(1-\hat{p}))$
2. Calculate the variance, $\hat{v}_{\hat{L}}$, either by linearization ($\hat{v}_{\hat{L}} \approx \hat{v}_{\hat{p}} \times (1/(\hat{p}(1-\hat{p})))^2$) or by replicated methods such as half sample or jackknife
3. Calculate the normal confidence interval for $\hat{L}$: $\hat{L} \pm z_\alpha \sqrt{\hat{v}_{\hat{L}}}$
4. Transform this interval back into the domain of $\hat{p}$ using the inverse transform $\hat{p} = 1/(1+\exp(-\hat{L}))$

An advantage of this approach is that it avoids the possibility that the confidence interval will lie outside the interval [0, 1]. Of course there is no guarantee that these confidence intervals have or nearly have the correct coverage probabilities. Korn and Graubard (1998) introduced another method based on exact binomial intervals which gave better coverage properties than the logit intervals in a simulation experiment.

Suppose we had a binomial situation, which would almost be the case if we had simple random sampling and a low sampling fraction. Then it is well-known that when there are x successes from n trials the Clopper-Pearson $(1-2\alpha)\%$ confidence interval $(p_l(x,n), p_u(x,n))$ can be expressed as (see Johnson et al. 1993, p. 130):

$$\begin{aligned} p_l(x,n) &= \frac{\nu_1 F(\nu_1, \nu_2, \alpha)}{\nu_2 + \nu_1 F(\nu_1, \nu_2, \alpha)} \\ p_u(x,n) &= \frac{\nu_3 F(\nu_3, \nu_4, \alpha)}{\nu_4 + \nu_3 F(\nu_3, \nu_4, \alpha)} \end{aligned} \tag{1}$$

where $\nu_1 = 2x,\ \nu_2 = 2(n-x+1),\ \nu_3 = 2(x+1),\ \nu_4 = 2(n-x)$, and $F(\nu_1, \nu_2, y)$ is the F distribution with v_1 and v_2 degrees of freedom.

The modification that Korn and Graubard suggest is to replace the sample size n in Equation (1) by the estimated *effective* sample size, i.e., the sample size divided by the estimated design effect (deff) for the proportion.[3] The number of successes, x, is then given by the effective sample size times the estimated proportion $\hat{p}$.[4]

Korn and Graubard examined the coverage probabilities of such intervals in a simulation experiment where the sample design was typical of household survey designs

[3] The design effect of an estimate is the ratio of the variance of the estimate under a complex design to the variance under simple random sampling. The estimate of the variance of $\hat{p}$ might use analytically derived estimators. However, it is common now to produce such variance estimates using replicated methods such as the balanced half sample method or the jackknife method, since these more easily accommodate nonlinear estimators of p which arise from poststratification or more general regression estimators.

[4] Having completed this work, we became aware of work which shows that the Clopper-Pearson confidence intervals are conservative in estimating binomial conference intervals for small sample sizes. The appendix discusses another method of estimating binomial confidence intervals and presents the results of applying the Korn and Graubard modification.

in the U.S.A. Such designs have a small number of clusters or psu's which themselves contain a large number of secondary sampling units (ssu's). A reasonably large number of ssu's are selected from each psu. Within an ssu there may be further stages of selection depending on what is the ultimate sampling unit (household or individual) and on cost. Generally for the sample design Korn and Graubard considered, the ratio between the smallest and largest sample design weights[5] is bounded by 10.

The results of their study on these types of cluster designs showed that the coverage properties of these exact deff-adjusted binomial confidence intervals had coverage probabilities closer to the nominal value than logit transform confidence intervals, normal based confidence intervals and intervals and intervals based on the Poisson approximation to the binomial.

However, the 1999 NZGS design, as described in Section 2, has a large number of small psus (households) with only one ssu selected (people), in marked contrast to the sort of design which Korn and Graubard used for their simulation study. The other major difference from the Korn and Graubard study is that the choice of estimator, and differential nonresponse, produced estimation weights which had a factor of 40 in the ratio of the largest to the smallest. Therefore we felt that before we used the Korn and Graubard method on the 1999 NZGS some further analysis using bootstrapping was required.

4. Bootstrap for Sample Surveys

The more complex the design the more difficult it is to find a resampling method whose bootstrap distribution is a good estimator of the real distribution. However, if one is interested in variance or mean squared error estimation then the method can be applied to a range of standard complex designs: see Rao and Wu (1998).

For a stratified simple random sample without replacement (SRSWOR) design Rao and Wu suggest a rescaling procedure which matches the analytic formula for linear statistics. If the stratum population sizes are large, and if the bootstrap sample is a simple random sample with replacement (SRSWR) from each stratum of the same size as the stratum sample then this procedure reduces to the naive bootstrap.

Davison and Hinkley (1997, p. 92) also discuss finite population sampling, outlining the use of the bootstrap for SRSWOR and for SRSWR, and then extend these results to stratified sampling with and without replacement. Sampling without replacement adds some complexities which adjust for finite population corrections, but for SRSWR the bootstrap method is not affected and for stratified random sampling with replacement, the bootstrap for SRSWR sampling can be applied within strata (Haslett and Wear 1985).

Sitter (1992) outlines three bootstrap methods for survey data: the rescaling method, the mirror-match method and the without replacement bootstrap. All of these methods, while able to be used on without replacement sampling schemes, assume exchangeability over the whole sample and population of some functions of $\{y_i, \pi_i\}$, where y_i is the value of the variable y for the ith unit in the population and π_i is the inclusion probability for the ith unit in the population. That is, distinct exchangeable groups within the sample and the

[5] Given no nonresponse and assuming the Horwitz-Thompson estimator, these weights are just the inverse of the probability of selection for the ultimate sampling unit.

population are not permitted, and the inclusion probabilities, and hence weights, for a given unit i in the bootstrap sample are fixed given i.

Below a more general alternative is proposed where, in general, sampling may be with or without replacement for both the original sample and for the bootstrap, where there may be more than one exchangeable group in the population where, in the formation of the bootstrap estimate, a weight $\{\pi_j^{-1}\}$, for unit $j \neq i$ may be applied to unit i when i and j both belong to the same exchangeable group.

Properties of this extended bootstrap include unbiasedness and consistency under fairly general conditions, as shown in Section 5 below.

5. Theory of the Exchangeable Bootstrap

The central idea behind the exchangeable bootstrap is to form the survey data into groups within which the survey variable is in some sense equivalent (or more formally exchangeable). The survey estimator is then used, but instead of using the survey response for a particular respondent with its designated weight, a random choice of response is made from its corresponding group and from the weights and bootstrap responses, a bootstrap estimate is formed in the usual way (once a draw has been made in this way for all sampled units). A simple exposition of how to implement this is given in Section 7.2.

More formally, let a design unbiased estimator $\hat{Y}$ of a total Y based on the original sample s be

$$\hat{Y} = \sum_{i \in s} \frac{Y_i}{\pi_i} \tag{2}$$

Note that in this case the exchangeable bootstrap estimators are formed as

$$\hat{Y}^* = \sum_{i \in s} \frac{Y_i^*}{\pi_i} \tag{3}$$

where for a given unit i in the sample s, Y_i^* is drawn from an SRSWR from the exchangeable group to which unit i belongs. An extension to SRSWOR with exchangeable groups following Davison and Hinkley (1997) is also possible. Equations (2) and (3) are applicable for both with and without replacement sampling, as is discussed in detail by Haslett (1985); this formulation includes SRSWR, SRSWOR, stratified random sampling with and without replacement and the various forms of cluster sampling, for example, as special cases.

Now the defining property of the exchangeable group g is that for all units i belonging to g the superpopulation expectation $\mathscr{E}(Y_i)$ is y_g, the superpopulation mean for the gth group which while it depends on g is independent of i. That is,

$$\mathscr{E}(Y_i | i \in g) = y_g, \quad g = 1, 2, \ldots G$$

so that the superpopulation expectation of the bootstrap sample estimate is

$$\mathscr{E}(\hat{Y}^*) = \sum_{i \in s} \frac{\mathscr{E}(Y_i | i \in g)}{\pi_i}$$

$$= \sum_g \sum_{i \in s_g} \frac{y_g}{\pi_i} \quad \text{where } s = \bigcup_{g=1}^{G} s_g$$

and as usual $i \in s$ and $i \in s_g$ indicate the sampled members of the total finite population and the sampled members in group g of the finite population, respectively. Letting E denote the design expectation,

$$E\mathscr{E}(\hat{Y}^*) = \sum_g E\left(\sum_{i \in s_g} \frac{y_g}{\pi_i}\right)$$

$$= \sum_g \sum_{i \in g} \frac{y'_g}{\pi_i} \pi_i$$

where in general $y'_g = E(y_g) = \sum_{i \in g} y_i / N_g$, where $y_i = \mathscr{E}(Y_i)$ and N_g is the population size in the group g and is not a superpopulation quantity. As usual, $i \in g$ refers to all N_g members of the finite population in group g. Hence

$$E\mathscr{E}(\hat{Y}^*) = \sum_g \sum_{i \in g} y'_g$$

$$= \mathscr{E}(Y)$$

Thus $\hat{Y}^*$ is joint design-superpopulation unbiased for the population total since $\hat{Y} = \sum_{i \in s}(Y_i / \pi_i)$ is design unbiased.

The extension from totals to means and proportions is straightforward. More generally $\hat{Y}^* = \sum_{i \in s} w_{i(s)} Y_i^*$, where $w_{i(s)}, i = 1, 2, \ldots, N$ are the sample survey weights (which may depend on the sample s), is joint design-superpopulation unbiased for the population mean, total or proportion if and only if $\hat{Y} = \sum_{i \in s} w_{i(s)} Y_i$ is design unbiased.

Note that for a given sample s, the average value (with respect to the superpopulation) of the bootstrap estimate of Y is

$$\mathscr{E}(\hat{Y}^*) = \sum_g \sum_{i \in s_g} \frac{y_g}{\pi_i}$$

which can be unbiasedly estimated by

$$\hat{\mathscr{E}}(\hat{Y}^*) = \sum_g \sum_{i \in s_g} \frac{\hat{y}_g}{\pi_i}$$

where $\hat{y}_g = (1/n_g) \sum_{i \in s_g} Y_i$ and Y_i is interpreted as the realized value for each unit $i \in s$ and n_g is the number of members of group g in the sample.

That is, by

$$\hat{\mathscr{E}}(\hat{Y}^*) = \sum_g \sum_{i \in s_g} \frac{(1/n_g)\sum_{i \in s_g} Y_i}{\pi_i}$$

$$= \sum_g \frac{1}{n_g} \left(\sum_{i \in s_g} Y_i \right) \left(\sum_{i \in s_g} \frac{1}{\pi_i} \right)$$

$$= \sum_g \sum_{i \in s_g} \frac{Y_i}{\pi_g} \quad \text{where } \pi_g = n_g \left(\sum_{i \in s_g} \frac{1}{\pi_i} \right)^{-1}$$

$$\approx \sum_g \sum_{i \in s_g} \frac{Y_i}{\pi_i} \quad \text{since } \pi_g \text{ is the harmonic mean}$$

$$= \sum_{i \in s} \frac{Y_i}{\pi_i} = \hat{Y}$$

So that, given the *correct* exchangeable groups, for all $i \in s$,

$$\hat{\mathscr{E}}(\hat{Y}^*) \approx \hat{Y}$$

for a given sample s.

The design-superpopulation unbiasedness for $\hat{Y}^*$ thus has an important practical consequence, if $\hat{Y}$ is known to be design unbiased.

Since

$$\mathscr{E}(\hat{Y}^*) = \sum_{i \in s} \frac{y_g}{\pi_i}$$

and y_g can be estimated by $\hat{y}_g$, then if the $\{y_g\}$ are distinct and the exchangeable groups are wrongly chosen, there will be sample units misassigned to their correct group g, and the difference between the usual sample survey estimate and the average estimated from the bootstrap $\hat{Y} - \hat{\mathscr{E}}(\hat{Y}^*)$ will often be appreciable.

There is consequently a simple check on whether exchangeable groups have been suitably chosen. If $y_g \approx y_{g'}$ for two groups g and g', then assignment of unit $i \in s$ to g or g' is of no particular consequence. However, if the two groups g and g' are clearly not a single exchangeable group because they have different means, and some sample units are misassigned, then the bias term $\hat{Y} - \hat{\mathscr{E}}(\hat{Y}^*)$ will usually be appreciable relative to the confidence intervals for $\hat{\mathscr{E}}(\hat{Y}^*)$ from the bootstrap.

Further the bootstrap estimator $\hat{Y}^*$ is not only design-superpopulation unbiased if $\hat{Y}$ is design unbiased, but also is consistent for Y if and only if $\hat{Y}$ is consistent for Y. That is,

$$P(|\hat{Y}_n^* - Y| < \epsilon) \to 1 \quad \text{as} \quad n \to N \text{ (WOR) or as } n \to \infty \text{ (WR)}$$

if and only if

$$P(|\hat{Y}_n - Y| < \epsilon) \to 1 \quad \text{as } n \to N \text{ (WOR) or as } n \to \infty \text{ (WR)}$$

where $\hat{Y}_n^* = \hat{Y}^*$ for sample size n, and $\hat{Y}_n = \hat{Y}$ for sample size n. (The result follows directly because $\hat{y}_g$ is consistent for y'_g for each g, for both with and without replacement sampling.)

The extension of these results to linearizable nonlinear statistics is straightforward, with approximate linearization via Taylor series expansion also being a possibility.

6. Confidence Intervals

Two methods of producing confidence intervals for the exchangeable bootstrap were considered. First we considered the percentile method. That produced unsatisfactory results so we then considered the accelerated bias corrected percentile method.

6.1. Percentile method

Suppose we have some parameter of interest θ and an estimator $\hat{\theta}$ for it. Bending the usual notation slightly, let $\hat{\theta}$ also be the estimate of θ from the full sample. For each bootstrap sample s we calculate an estimate $\hat{\theta}_s$. From the N bootstrap samples, form the empirical cumulative distribution function

$$ECDF(t) = \#\{\hat{\theta}_s \leq t\}/N$$

of the estimator $\hat{\theta}$. For some α between 0 and .5, define

$$\hat{\theta}_L(\alpha) = ECDF^{-1}(\alpha), \quad \hat{\theta}_U(\alpha) = ECDF^{-1}(1-\alpha)$$

where $ECDF^{-1}$ is the inverse of the empirical cumulative distribution function. The percentile method uses

$$[\hat{\theta}_L(\alpha), \hat{\theta}_U(\alpha)]$$

as an approximate $1 - 2\alpha$ confidence interval for $\hat{\theta}$.

6.2. Bias correction and acceleration methods

A check on the adequacy of the percentile method is whether the median of the empirical distribution function is the value of the estimate obtained from the full sample, i.e., does $\#\{\hat{\theta}_s \leq \hat{\theta}\}/N = .5$. If it does not, Efron (see e.g., Efron and Tibshirani 1993) suggests making a bias correction to the percentile method.

Let Φ be the cumulative distribution function for the standard normal random variable. Define

$$z_0 = \Phi^{-1}(ECDF(\hat{\theta}))$$

The bias corrected percentile method (BC method) uses

$$[EDCF^{-1}(\Phi(2z_0 - z_\alpha)), EDCF^{-1}(\Phi(2z_0 + z_\alpha))]$$

as an approximate $1 - 2\alpha$ confidence interval for $\hat{\theta}$, where z_α is, as before, the $1 - \alpha$ quantile of the standard normal distribution, i.e., $\Phi(z_\alpha) = 1 - \alpha$. Clearly if $\#\{\hat{\theta}_s \leq \hat{\theta}\}/N = .5$ then the BC method reduces to the percentile method.

Essentially the BC method makes an adjustment for potential asymmetry in the percentile method interval. But it assumes that the standard error of the estimator is the same for all values of the true parameter. Efron (see e.g., Efron and Tibshirani 1993) suggests this variance stabilizing assumption can be checked and adjusted for through the *acceleration* constant which measures the rate of change of the standard error of the estimator with respect to the true parameter. This leads to what he calls the BC_a method.

Specifically, define

$$z[\alpha] = z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}$$

where a is the acceleration constant. Similarly define $z[1-\alpha]$. The bias corrected percentile acceleration method (BC_a method) uses

$$[EDCF^{-1}(\Phi(z[\alpha])),\ \ EDCF^{-1}(\Phi(z[1-\alpha]))]$$

as an approximate $1-2\alpha$ confidence interval for $\hat{\theta}$. Here z_0 is the same as in the BC method. The acceleration constant needs to be estimated. Since a and z_0 are functions of the bias, variance and skewness of $\hat{\theta}_s$, Shao and Tu (1995) and Efron and Tibshirani (1993) suggest an approximation can be obtained using

$$\hat{a} = \frac{1}{6}\frac{\sum_{i=1}^{N}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{\left(\sum_{i=1}^{N}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2\right)^{3/2}}$$

where $\hat{\theta}_{(i)}$ is the estimate of the parameter with the ith unit of the sample deleted, and $\hat{\theta}_{(.)} = \sum_{i=1}^{N}\hat{\theta}_{(i)}/N$. Intuitively, $\hat{a}$ is the (jackknife estimate of skewness)/6.

This estimator of a required no great extra computation on our part as we already had produced *jackknife weights* to produce the sample errors of the major estimates of the 1999 NZGS. A point to note is that in the case of a complex survey design, modifications to the jackknife have to be made to adjust for the design, e.g., in a cluster design the unit which is deleted is not the ultimate sampling unit but the primary sampling unit. Use of psu's removes much of the complexity from the implementation of the jackknife, so that $\hat{a}$ is largely unaffected by the design.

6.2.1. Coverage properties for the exchangeable bootstrap

Because we are using the BC_a method on the exchangeable bootstrap, the theory of Davison and Hinkley (1997, pp. 211–220) applies. That is, the coverage properties of the exchangeable bootstrap are the same as the coverage properties of the BC_a. Strictly speaking this theory ignores the finite population corrections usually necessary but not so in this study because of the specific sample design. If finite population corrections were required, then modifying the bootstrap samples within strata, as in Davison and Hinkley (1997 p. 93) suffices.

7. Results

In the case of the 1999 NZGS variance estimates obtained from treating the sampling as with replacement increased the estimated variance from that for treating the sampling as without replacement by less than 2% over a wide range of variables when using

conventional estimators. Since this effect is negligible, especially when translated into confidence intervals, bootstrap samples for the 1999 NZGS were drawn with replacement.

Results from two different methods of producing the bootstrap samples and two different methods of producing confidence intervals are presented here.

7.1. Naive bootstrap

The first method tried was the naive bootstrap. As mentioned in Section 4, given the sample design of the 1999 NZGS, this method is essentially the rescaling method (see Rao and Wu 1988 and Sitter 1992).

The method consists of selecting a bootstrap SRSWR sample within each stratum. The response factors to account for nonresponse were calculated for groups defined by stratum and week of interview. The bootstrap selection weights were adjusted by these response factors to produce the response weights for a household. The response weights were adjusted by the number of eligible people in the household and the number of telephone lines to get the person weight. Finally the person weight was calibrated by the raking ratio method using age by sex and ethnic group benchmarks. For this study 5,000 bootstrap samples were taken.

An example of the results obtained is given in Table 1. Here the variable is lifetime problem gambling prevalence based on a modified South Oaks Gambling screen. Estimates are given for the New Zealand population and broken down by sex, age, and ethnicity. The survey estimates are typically small to very small. Note that half the time the mean of the naive bootstrap estimates (called bootstrap estimate in the table) are more than 10% different from the survey estimate, and can be over 100% different. Note that the percentile confidence interval for the Māori estimate does not cover the survey estimate.

Table 1. Results from the naive bootstrap for the variable Lifetime problem gambling for selected subpopulation estimates. The values are expressed in percentages

(Sub) population	Survey estimate	Bootstrap estimate	Percentile confidence interval		BC_a confidence interval	
			low	high	low	high
New Zealand	1.9	2.1	1.7	2.5	1.4	2.1
Male	2.8	3.1	2.4	3.8	1.9	3.3
Female	1.1	1.1	0.8	1.5	0.7	1.3
18–24	2.1	2.2	1.1	3.4	1.0	3.2
25–34	3.2	3.7	2.5	5.0	1.7	3.8
35–44	2.5	2.1	1.3	3.0	1.9	3.5
45–54	1.2	1.2	0.8	1.8	0.6	1.6
55–64	1.5	2.2	1.3	3.3	0.7	1.7
65 +	0.5	0.6	0.3	1.0	0.1	0.8
European	1.3	1.4	1.1	1.8	0.9	1.5
Māori	3.6	2.0	0.9	3.2	4.0	4.4
Pacific Island	7.8	11.1	6.1	16.5	3.6	9.6
Asian	2.9	5.9	2.4	9.8	1.4	3.3
Other	0.8	0.4	0.0	1.9	0.0	2.1

The BC_a confidence intervals are different from the percentile ones, even when the bootstrap mean and survey estimate agree. The BC_a confidence interval for the Māori estimate still does not cover the survey estimate.

Similar results were observed with other variables such as current problem gambling, current and lifetime pathological gambling. Clearly, the naive bootstrap is unsatisfactory.

7.2. Exchangeable bootstrap

Since the naive bootstrap produced unsatisfactory results we next considered the exchangeable bootstrap. This method requires determining a set of *exchangeable* groups within the sample population. The obvious choice is to use some of the groups which from the design perspective are thought exchangeable. So one might first choose the sample strata. The standard household survey designs used by Statistics New Zealand usually have between 100 and 200 strata, so one could expect them to provide groups which are exchangeable. However, in the 1999 NZGS the stratification was very broad, with only 18 strata, so one would not expect much success using them as exchangeable groups. This lack of success ought to be obvious from examining the bias term $\hat{Y} - \mathscr{E}(\hat{Y}^*)$ as discussed in Section 5.

We chose to ignore these strata and used ethnic group, since for the variables of interest in this study, proportions of problem and pathological gamblers, previous studies have shown that ethnicity is a strong determiner of gambling behaviour.

Finally we chose to ignore the geographical stratification and used all of the groups (sex, age, and ethnic group) used in the calibration estimator, since for the purposes of nonresponse adjustment these were considered exchangeable groups.

Having chosen a set of exchangeable groups, one sorts the sample by exchangeable group. Within an exchangeable group the weights are retained in the order they appear in the original sample. The data to attach to these weights are chosen by an SRSWR within the exchangeable group, as illustrated below.

Original sample data	Final weight	Exchangeable group	Index	Bootstrap sample data
Y_1	w_1	1	1	Y_1^*
Y_2	w_2	1	2	Y_2^*
Y_3	w_3	1	3	Y_3^*
. . .	. . .	. . .	. . .	. . .
Y_j	w_j	k	1	Y_1^*
Y_{j+1}	w_{j+1}	k	2	Y_2^*
. . .	. . .	. . .	. . .	. . .

So the ith sample weight in the exchangeable group is always applied to the ith unit selected in the bootstrap via Y_i^*, but that Y_i^* could be the sample value for *any* unit in the same exchangeable group as unit i belongs to. The unusual feature here is that if a unit in the original sample is drawn (one or more times) in the bootstrap sample it will only exceptionally retain its original final weight. This is a consequence of the superpopulation

exchangeability requirement. It is also possible to draw from $\{Y_i^*\}$ in group g with unequal weights corresponding to the design, but if the members of the a group are truly exchangeable, this is unnecessary.

An example of the results obtained is given in Table 2. Again the variable is lifetime problem gambling prevalence, and estimates are given for the New Zealand population broken down by sex, age and ethnicity. Now, only about one third of the time the mean of the bootstrap estimates are more than 10% different from the survey estimate, and the maximum difference is less than 50%. Also, now the percentile confidence interval for the Māori estimate covers the survey estimate. The BC_a confidence intervals are generally close to the percentile ones, the more so when the bootstrap mean and survey estimate agree.

Table 2. Results from the exchangeable bootstrap for the variable Lifetime problem gambling for selected subpopulation estimates. The values are expressed in percentages

(Sub) population	Survey estimate	Bootstrap estimate	Percentile confidence interval		BC_a confidence interval	
			low	high	low	high
New Zealand	1.9	1.8	1.4	2.2	1.6	2.4
Male	2.8	2.6	1.9	3.3	2.3	3.7
Female	1.1	1.1	0.7	1.6	0.6	1.4
18–24	2.1	2.3	1.0	4.0	0.7	3.5
25–34	3.2	2.7	1.6	4.0	2.4	4.7
35–44	2.5	2.3	1.3	3.3	1.6	3.5
45–54	1.2	1.2	0.5	1.9	0.5	1.9
55–64	1.5	1.5	0.4	2.6	0.2	2.4
65 +	0.5	0.6	0.2	1.1	0.1	0.9
European	1.3	1.3	0.9	1.6	1.0	1.7
Māori	3.6	3.6	1.8	5.7	1.7	5.6
Pacific Island	7.8	6.6	2.4	11.5	3.9	13.1
Asian	2.9	2.5	0.4	9.8	0.8	6.2
Other	0.8	1.3	0.0	4.0	0.0	2.8

Similar results were observed with variables such as current problem gambling, current and lifetime pathological gambling. The exchangeable bootstrap appears to give satisfactory confidence intervals.

7.2.1. Comparison of confidence intervals

Table 3 compares the confidence intervals for the variable lifetime problem gambling for selected subpopulation estimates and for different methods. Clearly if the sample size is large, as for the New Zealand population, the normal confidence interval and the Korn and Graubard interval are almost the same. But as the sample sizes decrease the differences become marked.

The results for must subpopulations are very similar when the Korn and Graubard and the exchangeable bootstrap, when appropriately bias corrected and variance stabilized, are compared. A possible exception is the Pacific Island group, for which the exchangeable bootstrap is more conservative. However, because this subpopulation is small, and the

Table 3. Comparison of confidence intervals for the variable Lifetime problem gambling for selected subpopulation estimates. Normal refers to the confidence intervals based on the normal approximation. K & G refers to those proposed by Korn and Graudbard using exact binomial confidence intervals where the sample size is divided by the design effect of the estimate. Percentile refers to the percentile confidence interval from an exchangeable bootstrap using all the calibration groups as exchangeable groups. BC_a refers to the percentile confidence interval which has been bias corrected and variance stabilized. The values are expressed in percentages

(Sub) population	Normal		K & G		Percentile		BC_a	
	low	high	low	high	low	high	low	high
New Zealand	1.4	2.5	1.4	2.5	1.4	2.2	1.6	2.4
Male	1.9	3.8	2.0	3.9	1.9	3.3	2.3	3.7
Female	0.6	1.5	0.7	1.6	0.7	1.6	0.6	1.4
18–24	0.9	3.4	1.1	3.8	1.0	4.0	0.7	3.5
25–34	1.7	4.6	1.9	5.0	1.6	4.0	2.4	4.7
35–44	1.4	3.5	1.5	3.8	1.3	3.3	1.6	3.5
45–54	0.5	1.8	0.6	2.0	0.5	1.9	0.5	1.9
55–64	0.2	2.7	0.5	3.3	0.4	2.6	0.2	2.4
65 +	0.1	1.0	0.2	1.2	0.2	1.1	0.1	0.9
European	1.0	1.7	1.0	1.7	0.9	1.6	1.0	1.7
Māori	1.5	5.7	1.8	6.4	1.8	5.7	1.7	5.6
Pacific Island	0.1	15.5	2.0	19.5	2.4	11.5	3.9	13.1
Asian	0.0	5.7	0.7	7.4	0.4	5.5	0.8	6.2
Other	−0.8	2.5	0.0	4.7	0.0	4.0	0.0	2.8

proportion of problem and pathological gamblers is low, the number of Pacific Island problem and pathological gamblers in the sample is fewer than 10, despite the total sample size over all subpopulations being 6,452. This difference between the methods is consequently not definitive and may be due to the large design effects for the Pacific Island subpopulation in part due to nonresponse. Similar results were observed with variables such as current problem gambling, current and lifetime pathological gambling.

Small proportions usually imply small numbers of sample respondents having that characteristic. If the method of confidence interval formation for the estimated proportion is stable and reliable, then a change in the survey response data of (say) one individual should have little effect on the location and range of the estimated CI. Data "perturbation" was used in this way to check the stability of the exchangeable bootstrap using calibration cells as exchangeable groups by increasing (or decreasing) the number of problem and pathological gamblers in the sample by one.

Confidence intervals for proportion estimates for the variables current problem, current pathological, lifetime problem, and lifetime pathological were reproduced nationally for the national population and age group, sex and ethnic group subpopulation using two perturbation methods:

1. add one problem/pathological gambler: one randomly selected individual in the target population found not to be a problem or pathological gambler is now deemed to be one;
2. remove one problem/pathological gambler: one randomly selected individual in the target population found to be a problem or pathological now deemed not to be one.

Figure 1 displays the empirical distributions of lifetime problem gambling for the Māori subpopulation for the bootstrap distribution on the original sample and the two perturbed samples. The bootstrap distribution on the original sample is fairly symmetric in the middle but has a long right tail. There is little difference in the ranges of the perturbed distributions and the shapes are broadly similar.

Figure 2 displays the empirical distributions of lifetime problem gambling for the Pacific Island subpopulation for the bootstrap distribution on the original sample and the two perturbed samples. The bootstrap distribution on the original sample is not so

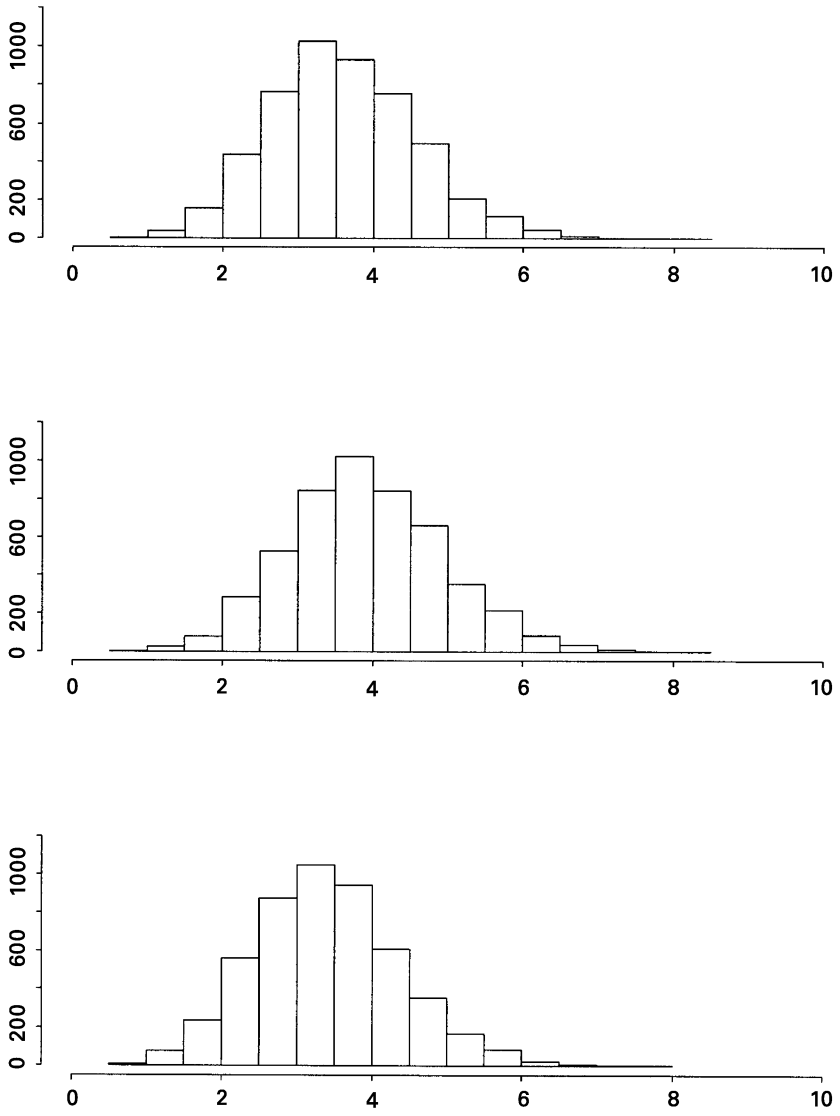

Fig. 1. Lifetime problem gambling for the Māori subpopulation. The top graph is the bootstrap distribution on the original sample. The middle graph is the bootstrap distribution for the original sample with one problem gambler added. The bottom graph is the bootstrap distribution for the original sample with one problem gambler removed. In all graphs the x-axis is the proportion expressed in percentages and the y-axis is frequency

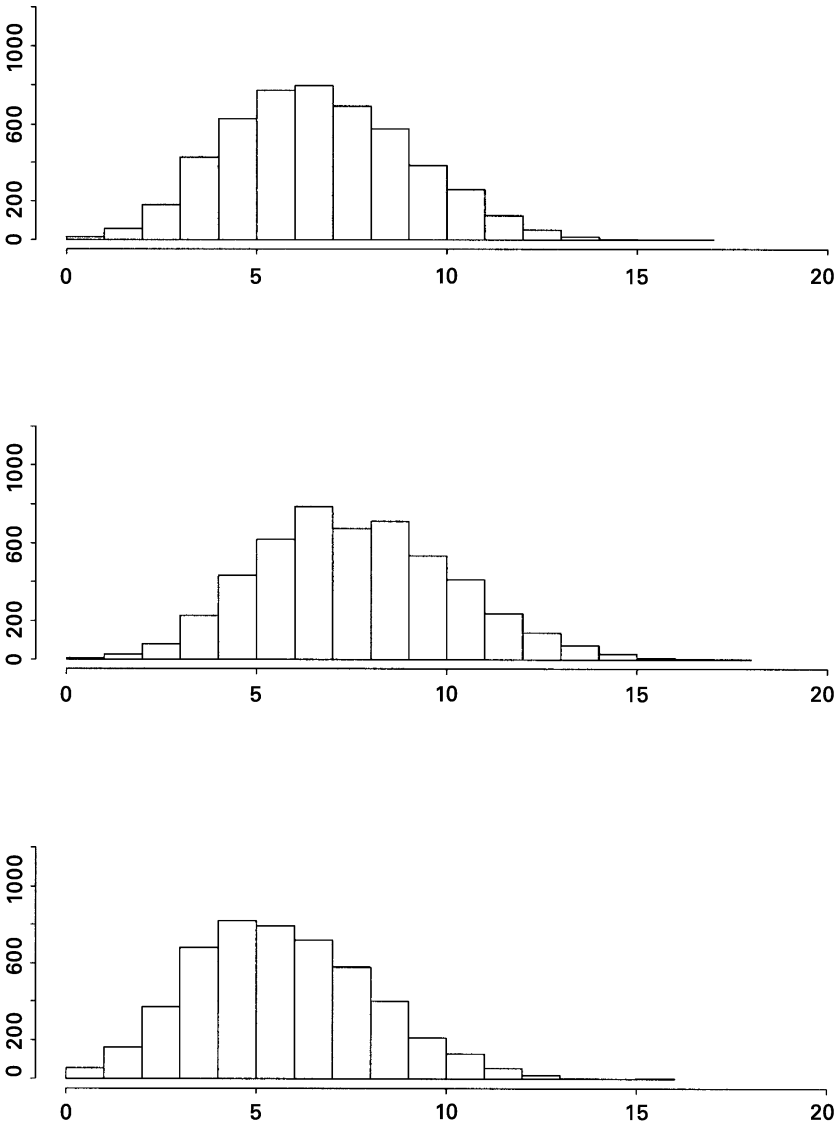

Fig. 2. Lifetime problem gambling for the Pacific Island subpopulation. The top graph is the bootstrap distribution on the original sample. The middle graph is the bootstrap distribution for the original sample with one problem gambler added. The bottom graph is the bootstrap distribution for the original sample with one problem gambler removed. In all graphs the x-axis is the proportion expressed in percentages and the y-axis is frequency

symmetric in the middle and has a longer right tail. But as with the Māori case the ranges of the perturbed distributions and the shapes are broadly similar.

Similar results were obtained for the other subpopulations and the other gambling variables. This is encouraging, suggesting that the exchangeable bootstrap is somewhat robust to small changes in the sample estimates.

8. Discussion and Conclusions

The Korn and Graubard method of creating confidence intervals has been applied to a real survey. The situation we were faced with in this survey was the need to provide a check on the Korn and Graubard method in a situation for which it was not developed, namely small proportions, small psu's and large variations in weights.

In the process of undertaking a bootstrap analysis, we have developed a new approach to bootstrap sampling for complex sample designs which looks promising when applied to real data. We have shown that it has desirable theoretical properties and for the gaming data provides estimates and coverage properties that are not highly sensitive to data changes.

The choice of exchangeable groups is not subjective. If an acceptable choice is made, the mean of the bootstrap estimates and the overall sample mean will match, otherwise not: thus the method provides a strong and useful diagnostic for the choice of groups.

We have found that the confidence intervals from the Korn and Graubard method and our exchangeable bootstrap for this data are similar. We are not arguing or able to argue for the primacy of either method, and like our analysis, even a simulation study would only have given a partial answer to this question. That said, this new approach is promising in itself and also suggests that the Korn and Graubard method is a reliable way of calculating confidence intervals for small proportions from a wider range of complex sample designs than that for which it was developed.

A. Appendix

A.1. Less Conservative Confidence Intervals

It is known that the Clopper-Pearson confidence intervals are conservative and a recent article by Brown, Cai and DasGupta (2001) suggests that the equal-tailed Jeffreys prior interval is a better alternative. This appendix compares the Clopper-Pearson and Jeffreys intervals for the gaming data using the exchangeable bootstrap as the benchmark.

Suppose as before we have a binomial situation and there are x successes from n trials, then the Jeffreys $(1 - 2\alpha)\%$ confidence interval $(p_l(x, n), p_u(x, n))$ can be expressed as:

$$\begin{aligned} p_l(0, n) &= 0 \\ p_u(1, n) &= 1 \\ &\text{and otherwise} \\ p_l(x, n) &= B(x + 1/2, n - x + 1/2, \alpha) \\ p_u(x, n) &= B(x + 1/2, n - x + 1/2, 1 - \alpha) \end{aligned} \tag{4}$$

where $B(\nu_1, \nu_2, y)$ is the Beta distribution with shape parameters ν_1 and ν_2.

To account for a complex sample design one would follow Korn and Graubard and replace the sample size n in Equation (4) by the estimated *effective* sample size. As before, the number of successes, x, is given by the effective sample size times the estimated proportion $\hat{p}$.

Table 4 shows the Clopper-Pearson and Jeffreys intervals both modified in the manner suggested by Korn and Graubard along with the intervals from the exchangeable bootstrap for the variable lifetime problem gambling. As expected, the Jeffreys intervals are typically shorter; more so on the right. For that reason they are closer to the intervals from the exchangeable bootstrap.

Table 4. Comparison of confidence intervals for the variable Lifetime problem gambling for selected subpopulation estimates. K & G refers to those proposed by Korn and Graubard using Clopper-Pearson binomial confidence intervals where the sample size is divided by the design effect of the estimate. Jeffreys refers to the equal-tailed Jeffreys prior interval with the modification proposed by Korn and Graubard. Percentile refers to the percentile confidence interval from an exchangeable bootstrap using all the calibration groups as exchangeable groups. BC_a refers to the percentile confidence interval which has been bias corrected and variance stabilized. The values are expressed in percentages

(Sub) population	K & G		Jeffreys		Percentile		BC_a	
	low	high	low	high	low	high	low	high
New Zealand	1.4	2.5	1.4	2.5	1.4	2.2	1.6	2.4
Male	2.0	3.9	2.0	3.9	1.9	3.3	2.3	3.7
Female	0.7	1.6	0.7	1.5	0.7	1.6	0.6	1.4
18–24	1.1	3.8	1.1	3.7	1.0	4.0	0.7	3.5
25–34	1.9	5.0	1.9	4.8	1.6	4.0	2.4	4.7
35–44	1.5	3.8	1.5	3.7	1.3	3.3	1.6	3.5
45–54	0.6	2.0	0.6	1.9	0.5	1.9	0.5	1.9
55–64	0.5	3.3	0.6	3.1	0.4	2.6	0.2	2.4
65 +	0.2	1.2	0.2	1.1	0.2	1.1	0.1	0.9
European	1.0	1.7	1.0	1.7	0.9	1.6	1.0	1.7
Māori	1.8	6.4	1.9	6.2	1.8	5.7	1.7	5.6
Pacific Island	2.0	19.5	2.6	18.0	2.4	11.5	3.9	13.1
Asian	0.7	7.4	0.9	6.8	0.4	5.5	0.8	6.2
Other	0.0	4.7	0.1	3.9	0.0	4.0	0.0	2.8

In practical terms, given nonresponse bias and measurement errors, use of the conservative Clopper-Pearson intervals with the Korn and Graubard modification is sensible in guarding against a Type I error. However, when nonresponse bias and measurement error is small, and control of Type II error is important, consideration should be given to using the Jeffreys interval.

9. References

Brown, L.D., Cai, T.T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16, 101–133.

Cochran, W.G. (1997). Sampling Techniques. (3rd edition). New York: John Wiley and Sons.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and Their Applications. New York: Cambridge University Press.

Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. New York: Chapman and Hall.

Haslett, S.J. (1985). The Linear Non-homogeneous Sample Survey Estimator. Sankhya, Series B, 47, 101–117.

Haslett, S.J. and Wear, R.G. (1985). Biomass Estimation of Artemia at Lake Grassmere, Marlborough, New Zealand. Australian Journal of Marine and Freshwater Research, 36, 537–557.

Johnson, N.L., Kotz, S., and Kemp, A.W. (1993). Univariate Discrete Distributions. (2nd edition). New York: John Wiley and Sons.

Kish, L. (1965). Survey Sampling. New York: John Wiley and Sons, Inc.

Korn, E.L. and Graubard, B.I. (1998). Confidence Intervals for Proportions with Very Small Expected Number of Positive Counts Estimated from Survey Data. Survey Methodology, 24, 193–201.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data. Journal of the American Statistical Association, 83, 231–241.

Rust, K.F. and Rao, J.N.K. (1996). Variance Estimation Techniques for Complex Surveys Using Replication Techniques. Statistical Methods in Medical Research, 5, 283–310.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer.

Sitter, R.R. (1992). Comparing Three Bootstrap Methods for Survey Data. Canadian Journal of Statistics, 20, 135–154.

Received January 2002
Revised July 2004