Survey Methodology, December 1987
Vol. 13, No. 2, pp. 135–146
Statistics Canada

Estimates Based on Randomly Rounded Data

C.S. WITHERS[1]

ABSTRACT

Methods are given to estimate functions of the cell probabilities associated with a table of multinomial data that has been randomly rounded to multiples of a given number, say l. We show that: (i) random rounding causes only second order effects on bias and variance; (ii) the loss of efficiency in using the natural estimates of cell probability is negligible provided that the cell entry is large compared with $(l^2 - 1)/(6R)$ where R is the number of cells in the table; and (iii) estimates of apparently exponentially small bias are available for moments of these natural estimates and for polynomials in the cell probabilities.

KEY WORDS: Random rounding; Bias reduction; Efficiency.

1. INTRODUCTION AND SUMMARY

This paper gives methods of estimating a function of the cell probabilities associated with a table of multinomial data that has been randomly rounded. Random rounding is a widely used method for preserving confidentiality in situations where an entry of 1 in a table might identify an individual and so break a confidentiality requirement. Instead of tabling the value of a table entry, say N, one rounds N to the nearest multiple of a given number l above N with probability (w.p.) α or below N w.p. $1 - \alpha$, where α is chosen so that the rounded value M satisfies

$$E(M \mid N) = N.$$

That is, if for some integer j, $jl \leq N < (j + 1)l$, then

$$M = \begin{cases} jl \text{ w.p. } 1 - \alpha \\ (j + 1)l \text{ w.p. } \alpha \end{cases} \tag{1.1}$$

where $\alpha = r/l$ and $r = N - jl$.

The rounding base l used by the Department of Statistics in New Zealand is $l = 3$, while Statistics Canada reportedly uses $l = 5$. See Penny and Ryan (1986).

Random rounding should not be confused with grouping or non-random rounding of sample values to the nearest integral multiple of l (associated with Sheppard's corrections for moments). Nor should it be confused with intentional contamination, another method of preserving confidentiality where one simply adds to N an independent random variable with mean 0. (The main disadvantage of intentional contamination is the possibility of a negative cell entry). For some references on these methods see Gastwirth *et al.* (1978) and Kendall and Stuart (1977). Some references on random rounding for multivariate data and grouped data are also given in Gastwirth *et al.* (1978).

[1] C.S. Withers, Applied Mathematics Division, Department of Scientific and Industrial Research, Box 1335, Wellington, New Zealand.

In this paper we confine our attention to problems of estimating a function of the cell probabilities associated with a table of R values that have been randomly rounded. For convenience we label these cell probabilities as $p_1, \ldots, p_R$ rather than $\{p_{ij}, 1 \le i \le I, 1 \le j \le J\}$, as is more usual for an $I \times J$ table.

Thus, $1 = \Sigma_1^R p_i$ and $n = \Sigma_1^R N_i$ is the sum of the entries in the table. Let $\{M_i\}$ be the rounded values of $\{N_i\}$. Given n, we assume $\{N_i\}$ has the multinomial distribution with parameters n and $\{p_i\}$. This is true with $p_i = m_i / \Sigma_j m_j$ if, unconditionally, $\{N_i\}$ are independent Poisson variables with means $\{m_i\}$.

Two unbiased estimates of p_1 are

$$p_1^* = N_1 / n \text{ and } \hat{p}_1 = M_1 / n. \tag{1.2}$$

The first is not a true estimate since N_1 is not made available. The second is the natural estimate. (We assume n is reported. If it is not, there is negligible difference in replacing n by $\Sigma_1^R M_i$.) However, other unbiased estimates exist, namely the "complementary estimate"

$$\tilde{p}_1 = -\sum_{j \ne 1} M_j / n, \tag{1.3}$$

and hence

$$p_1(\lambda) = (1 - \lambda)\hat{p}_1 + \lambda \tilde{p}_1 \text{ for any given } \lambda. \tag{1.4}$$

This raises the issue of what is the best λ to use, and what loss of efficiency there is in sticking to the natural estimate — that is, using $\lambda = 0$. An answer requires the variances of these estimators. These are given by

Theorem 1.1.

$$\operatorname{var}(\hat{p}_1) = (p_1 - p_1^2)\, n^{-1} + \{(l^2 - 1)/6 + \Delta_n(p_1)\} n^{-2} = v_n(p_1), \tag{1.5}$$

where

$$\Delta_n(p_1) = \sum_{i=0}^{l-1} i(l - i)\{P(N_1 \bmod l = i) - l^{-1}\}. \tag{1.6}$$

Also,

$$\operatorname{var}(\tilde{p}_1) = (p_1 - p_1^2) n^{-1} + \{(R - 1)(l^2 - 1)/6 + \sum_{j \ne 1} \Delta_n(p_j)\} n^{-2}, \tag{1.7}$$

and

$$\operatorname{var}(p_1(\lambda)) = (p_1 - p_1^2) n^{-1} + \{\alpha(\lambda)(l^2 - 1)/6 + \nabla_n(p)\} n^{-2}, \tag{1.8}$$

where

$$\alpha(\lambda) = (1 - \lambda)^2 + (R - 1)\lambda^2 \tag{1.9}$$

and

$$\nabla_n(p) = (1 - \lambda)^2 \Delta_n(p_1) + \lambda^2 \Sigma_{i \ne 1} \Delta_n(p_i). \tag{1.10}$$

Proofs of the theorems in this paper are given in Section 2.

In Appendix A we give evidence that for $0 < p_1 < 1$, $P(N_1 \bmod l = i) - l^{-1} \to 0$ exponentially fast as $n \to \infty$, so that $\Delta_n(p_1) \to 0$ exponentially fast as $n \to \infty$, and hence $\nabla_n(p)$ also, provided $p_i \neq 0$ for all i.

Since $\alpha(\lambda)$ is minimised by $\lambda_R = R^{-1}$ and $\alpha(\lambda_R) = 1 - R^{-1}$ so, asymptotically, is $\text{var}(p_1(\lambda))$. Hence the loss of efficiency in using the natural estimate $\hat{p}_1$ rather than the asymptotically optimal unbiased estimate $p_1(\lambda_R)$ when R is large, is

$$\{\text{var}(\hat{p}_1) - \text{var}(p_1(\lambda_R))\}/\text{var}(p_1(\lambda_R)) \approx (l^2 - 1)/\{6Rn(p_1 - p_1^2)\} \qquad (1.11)$$

which is negligible provided $M_1(1 - M_1/n) \approx n(p_1 - p_1^2)$ is large compared with $(l^2 - 1)/\{6R\}$.

Generally $M_1(1 - M_1/n)$ can be approximated by M_1. This then gives a convenient rule of thumb as to when the natural estimates are efficient. (If one or more $\{p_i\}$ are zero, since $p_i = 0$ implies $N_i = M_i = 0$, $\Sigma_{i \neq 1}$ must be interpreted as excluding cells for which $p_i = 0$, and R as the number of cells in the table for which $p_i \neq 0$.)

Using (1.5) we can now make a brief comparison with the method of contamination. The Australian and U.K. statistics departments reportedly round by adding to each cell entry 1 w.p. 1/4, 0 w.p. 1/2 and -1 w.p. 1/4, so that

$$\text{var}(\hat{p}_1) = (p_1 - p_1^2)n^{-1} + 1/2n^{-2}.$$

The factor 1/2 improves on 4/3 for the New Zealand system ($l = 3$) and 4 for the Canadian system ($l = 5$). The cost is less protection (a maximum change of 1 as opposed to 2 for the New Zealand system and 4 for the Canadian system), and a possibly negative cell entry if the procedure is applied to cells with zero entries.

Theorem 1.1 shows that random rounding has only a second order effect on the efficiency of estimating p_1 — the variance is only increased by a term of magnitude n^{-2}. The next result shows that this very important result is also true for estimating any smooth function of $\{p_i\}$. Set $r = R - 1$, $\mathbf{p} = (p_1, \ldots, p_r)$, $\mathbf{N} = (N_1, \ldots, N_r)$, $\mathbf{M} = (M_1, \ldots, M_r)$, $\mathbf{p}^* = \mathbf{N}/n$ and $\hat{\mathbf{p}} = \mathbf{M}/n$. Thus we have $\text{cov}(\mathbf{p}^*) = V/n$ where $V = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}'$. Suppose now we wish to estimate $f(\mathbf{p})$, a function with continuous second derivatives.

That is, $\dot{f}(\mathbf{p}) = \partial f(\mathbf{p})/\partial \mathbf{p}$ is a continuous $r \times 1$ function and $\ddot{f}(\mathbf{p}) = \partial^2 f(\mathbf{p})/\partial \mathbf{p} \partial \mathbf{p}'$ is a continuous $r \times r$ function.

Theorem 1.2. As $n \to \infty$ both $E(f(\mathbf{p}^*))$ and $E(f(\hat{\mathbf{p}}))$ equal

$$f(\mathbf{p}) + B(\mathbf{p})n^{-1} + O(n^{-2}) \text{ where } B(\mathbf{p}) = \text{trace } (\ddot{f}(\mathbf{p})V/2). \qquad (1.12)$$

Also both $\text{var}(f(\mathbf{p}^*))$ and $\text{var } (f(\hat{\mathbf{p}}))$ equal

$$v(\mathbf{p})n^{-1} + O(n^{-2}) \text{ where } v(\mathbf{p}) = \dot{f}(\mathbf{p})' V \dot{f}(\mathbf{p}). \qquad (1.13)$$

This theorem shows that

(a) random-rounding increases the variance of the natural estimate for $f(\mathbf{p})$ by only $O(n^{-2})$; and

(b) random-rounding likewise has only a second order effect on the bias of the natural estimate for $f(\mathbf{p})$.

According to (1.12), the natural estimate of $f(\mathbf{p})$, $f(\hat{\mathbf{p}})$, has bias of magnitude n^{-1}. We now show how to reduce this to n^{-2}.

Corollary 1.1. If for some function $f_n(\mathbf{p})$, $E(f_n(\mathbf{p}^*)) = f(\mathbf{p}) + O(n^{-2})$ then $E(f_n(\hat{\mathbf{p}})) = f(\mathbf{p}) + O(n^{-2})$.

Two such choices for $f_n(\hat{\mathbf{p}})$ are the "delta-estimate" for which

$$f_n(\mathbf{p}) = f(\mathbf{p}) - \left\{ \sum_{i=1}^{r} f_{ii}(\mathbf{p})p_i - \mathbf{p}'\ddot{f}(\mathbf{p})\mathbf{p} \right\} / (2n), \tag{1.14}$$

where $f_{ii}(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial p_i^2$, and the "jack-knife estimate" for which

$$f_n(\mathbf{p}) = nf(\mathbf{p}) - (n-1)\bar{f}, \tag{1.15}$$

where

$$\bar{f} = \Sigma_{i=1}^{r} p_i f([(n\mathbf{p} - e_i)/(n-1)]) + (1 - \Sigma_1^r p_i) f([n\mathbf{p}/(n-1)]),$$

$$e_i = \text{the } i\text{-th unit vector in } R^r,$$

$$\text{and } [x] : R^r \to R^r \text{ is defined by } [x]_i = \begin{cases} 0, & x_i < 0 \\ x_i, & 0 \le x_i \le 1. \\ 1, & x_i > 1 \end{cases}$$

These estimates were derived in Withers (1987a and 1987b). In particular, if $f(\mathbf{p})$ is only a function of p_1, say $f(\mathbf{p}) = g(p_1)$, then $f_n(\mathbf{p}) = g(p_1) - \ddot{g}(p_1)(p_1 - p_1^2)/(2n)$ and $\bar{f} = p_1 g([(np_1 - 1)/(n-1)]) + (1 - p_1) g([np_1/(n-1)])$. For example if $f(\mathbf{p}) = p_1^2$ then the delta-estimate uses $f_n(\mathbf{p}) = p_1^2\{1 - (1 - p_1)/n\}$.

We now illustrate that if $f(\mathbf{p})$ is a polynomial we can in fact find an estimate of $f(\mathbf{p})$ based on the natural estimate with bias apparently exponentially small. We do this for the case $f(\mathbf{p}) = p_1^2$.

Theorem 1.3. $\hat{\lambda}_1 = \{\hat{p}_1^2 - n^{-1}\hat{p}_1 - n^{-2}(l^2 - 1)/6\}(1 - n^{-1})^{-1}$ estimates $\lambda_1 = p_1^2$ with bias $\Delta_n(p_1)(n^2 - n)^{-1}$.

Similarly if $f_n(\mathbf{p})$ is a moment of $\hat{\mathbf{p}}$ then we can also find an estimate of $f_n(\mathbf{p})$ with bias apparently exponentially small. We illustrate this for the case $f_n(\mathbf{p}) = \text{var}(\hat{p}_1)$.

Theorem 1.4. $\hat{\lambda}_{2n} = n^{-1}(\hat{p}_1 - \hat{\lambda}_1) - n^{-2}(l^2 - 1)/6$ estimates $\lambda_{2n} = \text{var}(\hat{p}_1)$ with bias $-\Delta_n(p_1)(n^2 - n)^{-1}$.

These results may be generalised to higher order polynomials and moments using the expression for moments and cumulants of $\hat{\mathbf{p}}$ given in Appendix B. We now show that for the special case of $f(\mathbf{p})$ collinear, an unbiased estimate exists.

Theorem 1.5. Set $f_I(\mathbf{p}) = \Pi_{i=1}^{I} p_i$ where $1 \le I \le R$ and

$$a_{nI} = n^{-I} n!/(n-I)! = (1 - n^{-1})(1 - 2n^{-1}) \ldots (1 - \{I - 1\}n^{-1}). \tag{1.16}$$

Then

$$E(f_I(\hat{\mathbf{p}})) = E(f_I(\mathbf{p}^*)) = f_I(\mathbf{p}) a_{nI}. \tag{1.17}$$

Hence an unbiased estimate of $f(\mathbf{p})$ is $f_I(\hat{\mathbf{p}})/a_{nI}$.

Corollary 1.2. $\operatorname{cov}(\hat{p}_1,\hat{p}_2) = -p_1p_2/n$. Its unbiased estimate is $-\hat{p}_1\hat{p}_2/(n-1)$. More generally for $1 \le I \le R$, $E(\Pi_{i=1}^{I}(\hat{p}_i - p_i)) = c_{nI}\Pi_{i=1}^{I}p_i$ with unbiased estimate $(\Pi_{i=1}^{I}\hat{p}_i)\, a_{nI}/c_{nI}$ where $c_{nI} = \Sigma_{j=0}^{I}(-1)^{I-j}\binom{I}{j}a_{nj}$. (The same result holds with $\hat{\mathbf{p}}$ replaced by $\mathbf{p}^*$.)

From (1.16) one may derive unbiased estimates for other special polynominals in $\mathbf{p}$ such as p_1^2, $p_1p_2(p_1 + p_2)$ and $\Sigma_{i=1}^{R}p_i^3$ - but not for $p_1^2p_2$ or p_1^3.

Corollary 1.3. For $1 \le I < R$ an unbiased estimate of

$$f_I(\mathbf{p})\sum_{1}^{I} p_i \text{ is } f_I(\hat{\mathbf{p}})\left\{1 - In^{-1} - \sum_{I+1}^{R}\hat{p}_i\right\}/a_{n,I+1}. \tag{1.18}$$

In particular an unbiased estimate of p_1^2 is

$$\hat{p}_1(\hat{p}_1 - n^{-1})(1 - n^{-1})^{-1}. \tag{1.19}$$

We emphasize that the results of this paper are based on the assumption that table entries are independent Poisson's, or at least multinomial conditional on the total. The Poisson and multinomial models are appealing as they have a ready interpretation, and because sums of Poisson variables are Poisson. But sums of multinomials are multinomial only if they share the same cell probabilities $\mathbf{p}$. This suggests that conclusions drawn from such models may be less accurate if the populations modelled are composed of two or more inhomogeneous groups.

2. PROOFS

Proof of Theorem 2.1. Set $r = N_1 \bmod l$. Then (1.1) holds for $N = N_1$, $M = M_1$ with $jl = N - r$ and

$$E(M_1^2 \mid r) = (N_1 - r)^2(1 - r/l) + (N_1 - r + l)^2 r/l = N_1^2 + lr - r^2.$$

Hence

$$E(\hat{p}_1^2) = E(p_1^{*2}) + n^{-2}A_n(p_1), \tag{2.1}$$

where

$$A_n(p_1) = E(M_1^2 - N_1^2) = E(lr - r^2) = \sum_{i=0}^{l-1}(li - i^2)P(N = i)$$

$$= (l^2 - 1)/6 + \Delta_n(p_1)$$

since

$$l^{-1}\sum_{i=0}^{l-1} i(l - i) = (l^2 - 1)/6. \tag{2.2}$$

But

$$E(p_1^{*2}) = p_1^2 + (p_1 - p_1^2)n^{-1}, \tag{2.3}$$

so (1.5) follows. Now $\tilde{p}_1 = \hat{p}_1 - \sum (M_j - N_j)/n$,

$$\text{so } E(\tilde{p}_1^2) = E(\hat{p}_1^2) - 2n^{-2} \sum E(M_1(M_j - N_j)) + n^{-2} \sum E((M_i - N_i)(M_j - N_j))$$

$$= E(\hat{p}_1^2) - 2n^{-2}A_n(p_1) + n^{-2} \sum A_n(p_i)$$

since $E(\Pi_i f_i(M_i)|\{N_i\}) = \Pi_i E(f_i(M_i)|N_i)$. (2.4)

Hence $\text{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + n^{-2}\Sigma_{i\neq 1} A_n(p_i)$ so (1.7) holds.

Also,

$$E(\hat{p}_1\tilde{p}_1) = p_1 - n^{-2} \sum_{i\neq 1} E(M_1M_i) = p_1 - \sum_{i\neq 1} E(p_1^*p_i^*)$$

$$= p_1 - \sum_{i\neq 1} p_1p_i(1 - n^{-1}) = p_1 - p_1(1 - p_1)\ (1 - n^{-1}),$$

so

$$\text{cov}(\hat{p}_1,\tilde{p}_1) = (p_1 - p_1^2)n^{-1}. \tag{2.5}$$

Hence $\text{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{(1 - \lambda)^2A_n(p_1) + \lambda^2 \Sigma_{i\neq 1} A_n(p_i)\}n^{-2}$ and (1.8) holds.

Proof of Theorem 1.2. This was proved for $\mathbf{p}^*$ in Withers (1987a). Also since $\dot{f}$ is finite in a neighborhood of $\mathbf{p}$,

$$f(\hat{\mathbf{p}}) = f(\mathbf{p}^*) + (\hat{\mathbf{p}} - \mathbf{p}^*)'\dot{f}(\mathbf{p}^*) + O(|\hat{\mathbf{p}} - \mathbf{p}^*|^2).$$

$$E((\hat{\mathbf{p}} - \mathbf{p}^*)|N) = 0,\ E((\hat{p}_1 - p_i^*)^2|N) = 2n^{-2}I(N_1 \bmod l \neq 0),$$

where $I(A) = 1$ or 0 for A true or false, that is, $I(\cdot)$ is the indicator function Hence $E(f(\hat{\mathbf{p}})) = E(f(\mathbf{p}^*)) + O(n^{-2})$ and $\text{var}(f(\hat{\mathbf{p}})) = \text{var}(f(\mathbf{p}^*)) + O(n^{-2})$.

Proof of Theorem 1.3. This follows directly from (2.1) and (2.3).

Proof of Theorem 1.4. This follows from (2.1) and (1.5).

Proof of Theorem 1.5. The first equality in (1.16) follows from (2.4), and the second from the multinomial theorem. Corollary 1.2 follows immediately.

Proof of Corollary 1.3. From (1.16), for $1 \leq I < i \leq R$ we have

$$E(f_I(\hat{\mathbf{p}})\hat{p}_i) = f_I(\mathbf{p})p_ia_{n,I+1}$$

so

$$E(f_I(\hat{\mathbf{p}}) \sum_{I+1}^{R} \hat{p}_i / a_{n,I+1}) = f_I(\mathbf{p})(1 - \Sigma_1^I p_i)$$

$$= E(f_I(\hat{\mathbf{p}}) / a_{nI}) - f_I(\mathbf{p}) \Sigma_1^I p_i.$$

ACKNOWLEDGEMENT

I wish to thank Peter McGavin for doing the computing in Appendix A.

APPENDIX A

One expects that for f a smooth function

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) \to f(\mathbf{p})l^{-s} \tag{A.1}$$

as $n \to \infty$ provided $0 < p_i < 1$ for $1 \leq i \leq s \leq R$.

If $E(f(\hat{\mathbf{p}})) = f(\mathbf{p})$, one expects the rate of convergence to be exponential, $O(e^{-\lambda n})$ for some $\lambda > 0$. If $f(\hat{\mathbf{p}})$ is biased, then its bias is $O(n^{-1})$, so that one would expect this rate also to apply to (A.1). Convergence will in general break down as $\mathbf{p}$ approaches the boundary of $[0,1]^r$, since

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l))$$

$$= \begin{cases} f(\mathbf{p})I(j_1 = j_2 = \ldots = j_s = 0) & \text{if } \mathbf{p} = 0 \\ f(\mathbf{p})I(j_1 = n \bmod l) & \text{if } p_1 = 1. \end{cases}$$

To test these expectations we considered the case $s = 1$, $l = 3$, $j = 0$ and the functions (a) $f(\mathbf{p}) = 1$, (b) $f(\mathbf{p}) = p_1$, and (c) $f(\mathbf{p}) = \exp(p_1)$. Computations were done in quadruple precision on a VAX11/780, giving a precision for

$$\Delta = E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) - f(\mathbf{p})l^{-s}$$

of 112 bits - nearly 34 decimal places. Figures 1a, 1b and 1c plot Δ versus p_1 for $n = 6$, 18, 54. Since $n \bmod 3 = 0$, Δ is symmetric about $p_1 = 1/2$ for (a).

Since $\Delta = 2/3f(0)$ at $p_1 = 0$, and is equal to 2/3, 0 and 2/3 for (a), (b) and (c) respectively, convergence breaks down at $p_1 = 0$ for (a) and (c), but not for (b). At $n = 18$, Δ is already negligibly different from 0 for p_1 in (.2, .8) for (a) and for p_1 in (.1, .8) for (b) and (c). At $n = 54$, these ranges have grown to cover (.1, .9) for (a), (.02, .95) for (b), and (.07, .95) for (c).

Figures 2a and 2b plot $Y = \log(-\log|\Delta|)$ versus $X = \log(n)$ for (a) $f(\mathbf{p}) = 1$ and (b) $f(\mathbf{p}) = p_1$. As expected, except for small n, the curves are roughly parallel to $Y = X$ (except for (b) with $p_1 = .01$), consistent with $\Delta = O(e^{-\lambda n})$ for some $\lambda > 0$. The curves are not smooth, as Δ has only been calculated at n a power of 2 ($n = 2^i$ for $0 \leq i \leq 7$).

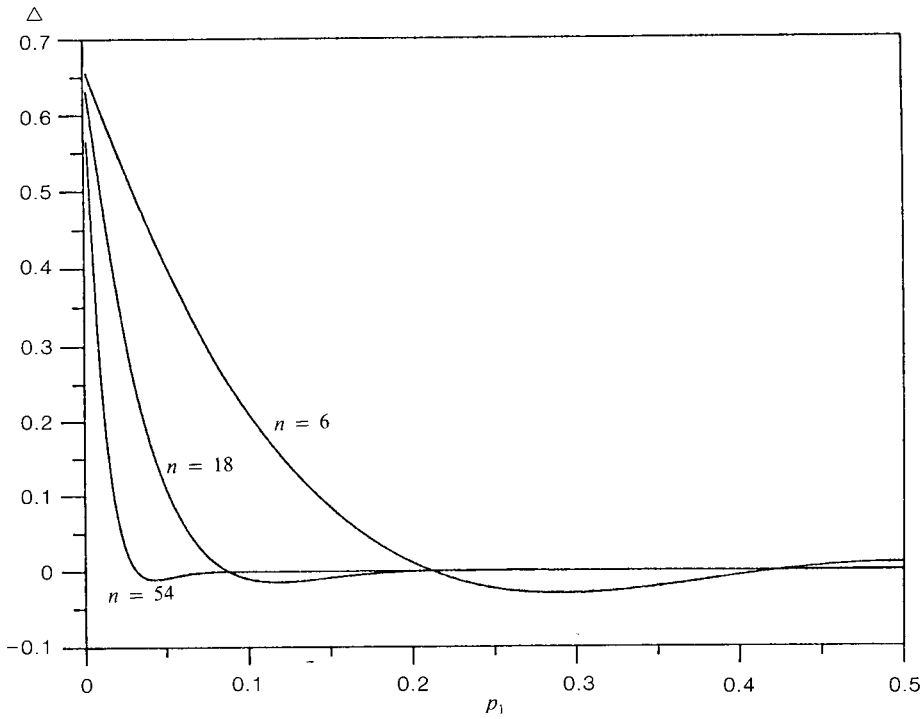

Figure 1a. Evidence for (A.1) When $f(\mathbf{p}) = 1$.

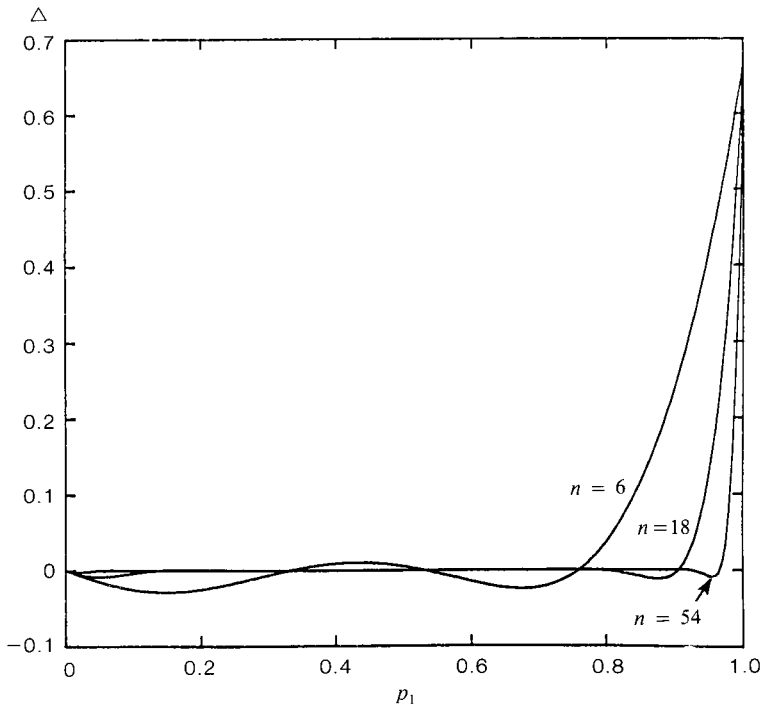

Figure 1b. Evidence for (A.1) When $f(\mathbf{p}) = p_1$.

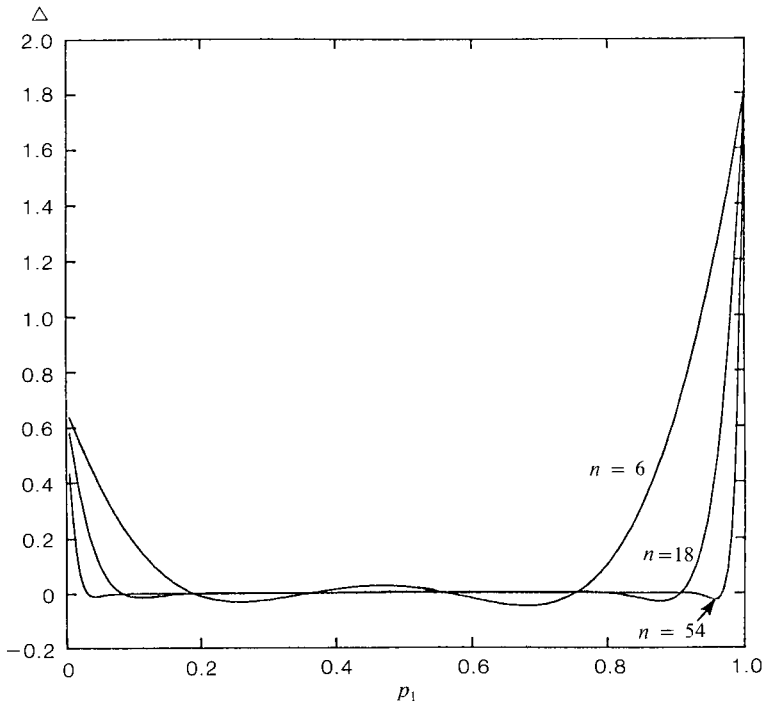

Figure 1c. Evidence for (A.1) When $f(\mathbf{p}) = \exp(p_1)$.

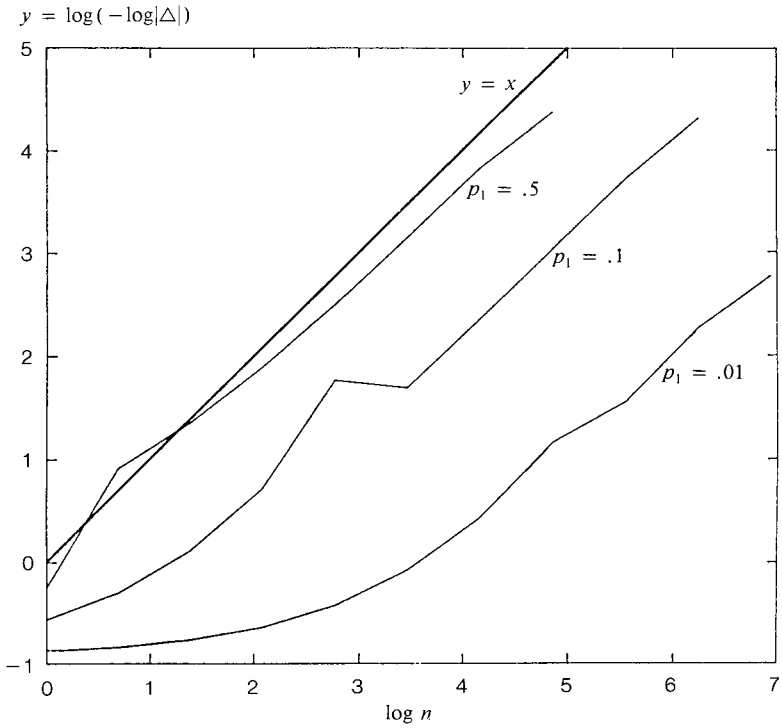

Figure 2a. Evidence for Exponential Convergence in (A.1) for $f(\mathbf{p}) = 1$.

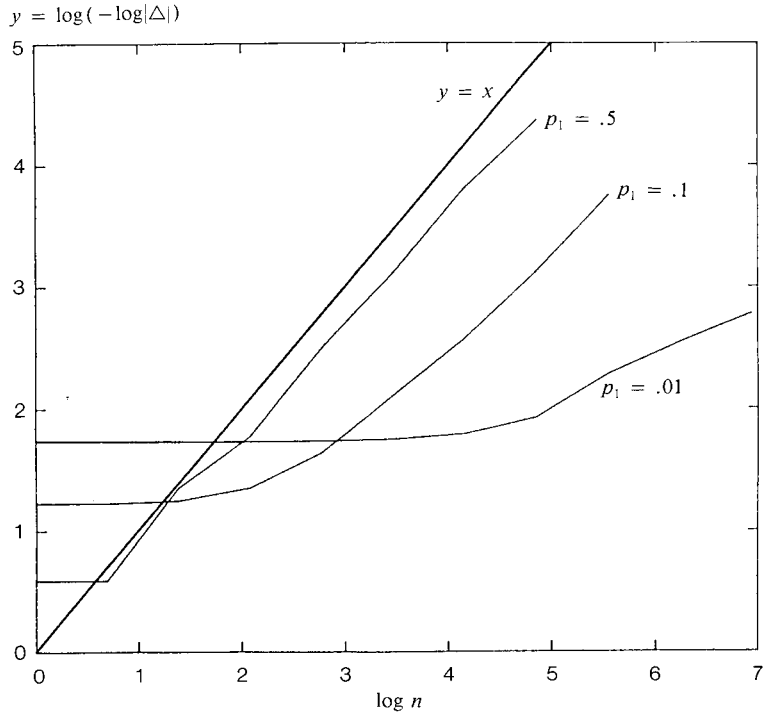

Figure 2b. Evidence for Exponential Convergence in (A.1) for $f(\mathbf{p}) = p_1$.

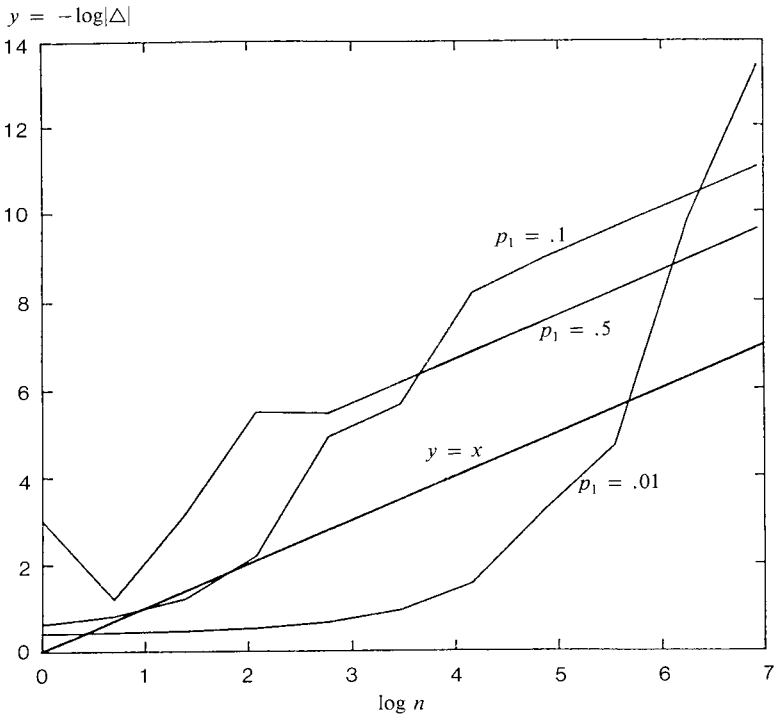

Figure 3. Evidence for Convergence at Rate $\sim n^{-1}$ in (A.1) for $f(\mathbf{p}) = \exp(p_1)$.

Figure 3 plots $Y = -\log \mid \Delta \mid$ versus $X = \log(n)$ for (c) $f(\mathbf{p}) = \exp(p_1)$. For n large the curves are parallel to $Y = X$ for $p_1 = .5$ and .1 consistent with $\Delta = O(n^{-1})$, but for $p_1 = 0.1$ the increase is much faster than linear. The graphs generally confirm our expectations on the rate of convergence in (A.1). To obtain analytic proofs would appear to require some sophisticated number theory.

APPENDIX B

Here we compare the moments and cumulants of $\mathbf{p}^* = N/n$ and $\hat{\mathbf{p}} = M/n$. Set $q_1 = 1 - p_1$, $n_i = N_i \bmod l$, and $m_i(j) = E(p_1^{*i} I(n_1 = j)) \to p_1^j / l$ as $n \to \infty$, assuming $p_1 \neq 0$ or 1. Elementary calculations yield

$$\mu(\hat{\mathbf{p}}) = \mu(\mathbf{p}^*) = \mathbf{p},$$

$$\mu_2(\hat{p}_1) = \mu_2(p_1^*) + M_{22} n^{-2} = p_1 q_1 n^{-1} + O(n^{-2}),$$

where

$$M_{22} = A_n(p_1) = \sum_{i=0}^{l-1} i(l - i) m_0(i) \to (l^2 - 1)/6$$

as $n \to \infty$,

$$\mu_3(\hat{p}_1) = \mu_3(p_1^*) + 3n^{-2} \sum_{j=0}^{l-1} (lj - j^2)\{m_2(j) - 2p_1 m_1(j) + p_1^2 m_0(j)\}$$

$$+ n^{-3} \sum_{j=0}^{l-1} a_{jl} m_0(j)$$

$$= \mu_3(p_1^*) + o(n^{-2}) = p_1 q_1 (1 - 2p_1) n^{-2} + o(n^{-2}),$$

and

$$a_{jl} = -j^3(1 - j/l) + (l - j)^3 j / l.$$

Similarly $\mu_4(\hat{p}_1)$ has the form $\mu_4(p_1^*) + \Sigma_2^4 M_{4i} n^{-i} = O(n^{-2})$ and $\kappa_4(\hat{p}_1)$ has the form $\Sigma_2^4 k_{4i} n^{-i}$ where $k_{42} = M_{42}$ does not converge to 0 as $n \to \infty$. Hence $\kappa_4(\hat{p}_1) \sim n^{-2}$, not n^{-3}. Hence $\hat{\mathbf{p}}$ does not satisfy the Cornish-Fisher assumption that $\kappa_r(\hat{\mathbf{p}}) = O(n^{1-r})$ for $r \geq 1$: see for example Kendall and Stuart (1977).

Moments and cumulants may also be obtained from the m.g.f. (moment generating function), which we now obtain.

$$E(\exp(t_1 M_1 / n) \mid N_1) = \exp(t_1 N_1 / n) S(t_1, n_1)$$

where

$$S(t_1, n_1) = (1 - n_1/l) \exp(-n_1 t_1 / n) + (n_1/l) \exp(l - n_1) t_1 / n.$$

Hence by (2.4), the m.g.f. is

$$E(\exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)) S(t) \text{ where } S(t) = \Pi_1^r S(t_i, n_i).$$

Also at $t = 0$, $S_1 = 0$ and so $S_{ij\ldots} = 0$ if a subscript occurs exactly once. For example, setting

$$S = S(t),\ \partial_i = \partial/\partial t_i,\ S_i = \partial_i S,\ S_{ij} = \partial_i \partial_j S,\ \ldots$$

gives

$$E(\hat{p}_1^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2} S + 2p_1^* S_1 + S_{11}\}),$$

$$E(\hat{p}_1^2 \hat{p}_2^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2}(p_2^{*2} S + 2p_2^* S_2 + S_{22}) + 2p_1^*(p_2^{*2} S_1 + 2p_2^* S_{12} + S_{122}) + (p_2^{*2} S_{11} + 2p_2^* S_{112} + S_{1122})\}).$$

Hence $E(\hat{p}_1^2) = E\{p_1^{*2} + S_{11}(0)\}$ and

$$E(\hat{p}_1^2 \hat{p}_2^2) = E\{p_1^{*2} p_2^{*2} + p_1^{*2} S_{22}(0) + p_2^{*2} S_{11}(0) + S_{1122}(0)\}.$$

where $S_{ii}(0) = S_{11}(0, n_i) = n^{-2}(l - n_i)n_i = n^{-2} \Sigma_{k=0}^{l-1} (l - k)kI(n_i = k)$ and $S_{1122}(0) = S_{11}(0)\, S_{22}(0)$. Some further simplifications can be obtained using $N_2 \mid N_1 \sim Bi(\theta, n - N_1)$ where $\theta = p_2/(1 - p_1)$. From the multinomial m.g.f. one obtains

$$E(p_1^{*2} p_2^{*2}) = n^{-4} p_1 p_2 \{(n)_4 p_1 p_2 + (n)_3 (p_1 + p_2) + (n)_2\}$$

where $(n)_i = n!/(n - i)! = n(n - i) \ldots (n - i + 1)$.

REFERENCES

GASTWIRTH, J.L., KRIEGE, A.M., and RUBIN, D.B. (1978). Statistical analyses from summary data and their impact on the issue of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 183-188.

KENDALL, M.G., and STUART, A. (1977). *The Advanced Theory of Statistics, Volume* 7. London: Griffin.

NARGUNDAR, M.S., and SAVELAND, W. (1972). Random rounding to prevent statistical disclosures. *Proceedings of the Social Statistics Section, American Statistical Association,* 382-385.

PENNY, R., and RYAN, M. (1986) A problem associated with random-rounding. *New Zealand Statistician,* 21, 43-52.

WITHERS, C.S. (1987a). Bias reduction by Taylor series. *Communications in Statistics - Theory and Methods* (forthcoming).

WITHERS, C.S. (1987b). Jackknifing binomials and multinomials. Unpublished manuscript, Department of Scientific and Industrial Research.