On a symptotic distributions in random sampling from finite populations

Paul Knottnerus

Discussion paper (09002)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2005-2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/’06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/’04–2005/’06	= crop year, financial year, etc. 2003/’04 to 2005/’06 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

Information
Telephone .. +31 88 570 70 70
Telefax .. +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax .. +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008309001 X-10

On asymptotic distributions in random sampling from finite populations

Paul Knottnerus

Summary

Existing proofs of central limit theorems in random sampling from finite populations are quite lengthy. The present paper shows how the proof for this kind of theorem in random sampling can be simplified by using the central limit theorem for independent, random vectors. Furthermore, use is made of the relationship between sampling without replacement and Poisson sampling given the condition that the sample size is n. The theorems deal with both equal and unequal probability sampling. In the latter case an approximation formula for the variance emerges as by-product from the central limit theorem without using second order inclusion probabilities.

Keywords: Central limit theorem; Error normal approximation; Simple random sampling; Unequal probability sampling; Variance estimation.

1. Introduction

Proofs of the central limit theorem for equal probability sampling from a finite population can be found in Madow (1948), Erdös and Renyi (1959), and Hájek (1960). For rejective Poisson sampling with varying probabilities a proof can be found in Hájek (1964). All proofs are technically difficult to demonstrate and omitted in most texts. Often simulations are used for demonstration; see Bellhouse (2001).

The main aim of this paper is to provide less intricate proofs for the central limit theorems in probability sampling from finite populations. Proofs in this paper are based on a generalization of the central limit theorem for a sequence of mutually independent two-dimensional random variables.

The outline of the paper is as follows. Section 2 examines the asymptotic behaviour of estimators in simple random sampling. Section 3 examines the asymptotic behaviour of estimators in unequal probability sampling. Unlike Hájek (1964) the author doesn't use (approximate) second order inclusion probabilities.

2. Asymptotic distributions in simple random sampling

Consider a population U of N numbers $U = \{y_1, \ldots, y_N\}$. Suppose that as $N \to \infty$ the population mean $\overline{Y}_N$ and the population variance σ^2_{Ny} converge to $\overline{Y}$ and σ^2_y, respectively. Let $\overline{y}_s$ denote the sample mean from a simple random sample of fixed size n $(= f_N N)$ without replacement (SRS). The sampling fraction f_N is such that $|f_N - f| < 1/N$ where f is a fixed number $(0 < f < 1)$.

Next, define for the Poisson sampling design with equal probabilities the following unbiased estimators for the population mean $\overline{Y}_N$ and n

$$\begin{aligned} \widehat{\overline{Y}}_{PO} &= \frac{1}{N}\sum_{k=1}^{N} a_k \frac{y_k}{f_N} \\ n_s &= \sum_{k=1}^{N} a_k, \end{aligned}$$

where the a_k are independent and identically distributed random variables. That is,

$$a_k = \begin{cases} 1 & \text{if the sample includes element } k \\ 0 & \text{otherwise} \end{cases}$$

with $P(a_k = 1) = f_N$ $(k = 1, \ldots, N)$.

In addition, define the random two-dimensional vector x_k and the matrix Σ_k by

$$\begin{aligned} x_k &= \begin{pmatrix} x_{1k} \\ x_{2k} \end{pmatrix} = a_k \begin{pmatrix} \frac{y_k}{f_N} \\ 1 \end{pmatrix} \\ \Sigma_k &= f_N(1 - f_N) \begin{pmatrix} \frac{y_k^2}{f_N^2} & \frac{y_k}{f_N} \\ \frac{y_k}{f_N} & 1 \end{pmatrix}, \end{aligned}$$

respectively. Σ_k is the covariance matrix of x_k. Other useful definitions and formulas in this context are

$$\begin{aligned} \overline{x}_N &= \frac{1}{N}\sum_{k=1}^{N} x_k = \begin{pmatrix} \widehat{\overline{Y}}_{PO} \\ n_s/N \end{pmatrix} \\ E(\overline{x}_N) &= \begin{pmatrix} \overline{Y}_N \\ f_N \end{pmatrix} \\ \overline{\Sigma} &= \lim_{N\to\infty} \overline{\Sigma}_N = \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} \Sigma_k \\ &= (1 - f) \begin{pmatrix} \frac{(\sigma_y^2 + \overline{Y}^2)}{f} & \overline{Y} \\ \overline{Y} & f \end{pmatrix} \end{aligned}$$

Denote the typical element ij of $\overline{\Sigma}$ by σ_{ij} and that of $\overline{\Sigma}^{-1}$ by σ^{ij} $(1 \leq i, j \leq 2)$. Applying the central limit theorem for vectors to the mutually independent x_k yields the following theorem.

Theorem 1. As $N \to \infty$ the distribution of $\sqrt{N}\{\overline{x}_N - E(\overline{x}_N)\}$ tends to the bivariate normal distribution with zero expectation and covariance matrix $\overline{\Sigma}$, provided that the data satisfy the Lindeberg condition.

For the Lindeberg condition, see Hájek (1964, page 1500) and Feller (1971, pages 262-3). Also note that the Lindeberg conditions for one dimension carry over to two-dimensional vectors due to the Cramér-Wold device; see Basu (2004, page 149).

In order to apply Theorem 1 to the problem of the limiting distribution of $\overline{y}_s$, define for an arbitrary constant u_0

$$\begin{aligned}
P_0 &= P(\overline{y}_s \leq \overline{Y}_N + u_0 h_1) \\
h_1^2 &= \frac{|\overline{\Sigma}|}{N\sigma_{22}} \quad (= \frac{1}{N\sigma^{11}} = (1-f)\frac{\sigma_y^2}{n}) \\
h_2^2 &= \frac{\sigma_{22}}{N} \quad (= \frac{f(1-f)}{N}) \\
\varphi(u) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2) \\
\Phi(u) &= \int_{-\infty}^{u} \varphi(x)dx \\
\psi(x;\Sigma) &= \frac{\exp(-x'\Sigma^{-1}x/2)}{2\pi\,|\Sigma|^{1/2}}
\end{aligned}$$

where in the last line $x = (x_1, x_2)'$ and $\Sigma =$cov(x). Recall from statistics theory that $1/\sigma^{11}$ is the conditional variance of x_1 given x_2. Furthermore, define λ_N by $\lambda_N = \mu_{3N}/\sigma_N^3$ and λ by $\lambda = \lambda_\infty$ where, given N, σ_N^2 and μ_{3N} stand for the variance and the third central moment of x_{2k}, respectively. Note that $\lambda_N = \lambda_N(f_N)$.

The following theorem specifies the error of the approximation of the lattice distribution of $\overline{x}_{2N}$ by the normal distribution.

Theorem 2. For a midpoint x of the lattice distribution F_{2N} of $\sqrt{N}\{\overline{x}_{2N} - E(\overline{x}_{2N})\}/\sigma_N$ it holds that as $N \to \infty$

$$F_{2N}(x;\lambda_N) - \Phi(x) = \omega_{2N}(x;\lambda) + o(1/\sqrt{N}) \tag{1}$$

where

$$\omega_{2N}(x;\lambda) = \frac{\lambda}{6\sqrt{N}}(1-x^2)\varphi(x) \tag{2}$$

Proof. For the lattice distribution F_{2N} it holds that for a midpoint x of the lattice

$$F_{2N}(x;\lambda_N) - \Phi(x) = \omega_{2N}(x;\lambda_N) + o(1/\sqrt{N}) \tag{3}$$

For a proof, see Feller (1971, page 540). Note that Feller assumes that $\lambda_N = \lambda$. However, since $\lambda_N = \lambda\{1 + o(1)\}$ the proof keeps valid when λ is replaced by λ_N. Moreover, since the derivatives of λ_N and ω_{2N} as functions of f_N are continuous at the point f and $|f_N - f| < 1/N$, it holds that

$$|\lambda_N - \lambda| \;=\; O(1/N)$$

$$|\omega_{2N}(x;\lambda_N) - \omega_{2N}(x;\lambda)| \;\; = \;\; O(1/N)$$

from which (1) follows. This concludes the proof.

Expression (2) for the error $\omega_{2N}(x)$ is based on a more refined Taylor approximation of the corresponding characteristic function $\chi_2(t/\sigma_N\sqrt{N})$ resulting in

$$\chi_{2rf}^N(t/\sigma_N\sqrt{N}) \sim (1 + \frac{\lambda_N(it)^3}{6\sqrt{N}}) \exp(-t^2/2)$$

instead of $\exp(-t^2/2)$. In addition, if F_{2N} is not a lattice distribution (3) is true for all x. Similar results can be derived for two-dimensional random vectors or when the x_{2k} have different variances for a given N; see Feller (1971, pages 521-548). Now the following theorem on the asymptotic behaviour of $\overline{y}_s$ can be proved. The notation "$A \sim B$" is used to indicate that A/B tends to unity as $N \to \infty$.

Theorem 3. As $N \to \infty$ it holds under the Lindeberg condition that $P_0 \sim \Phi(u_0)$.

Proof. Denote $\overline{Y}_N + u_0 h_1$ by m_0. It follows from the equivalence of SRS sampling and Poisson sampling conditional on a given n or f_N that

$$\begin{aligned} P_0 &= P(\overline{x}_{1N} \leq m_0 \,|\overline{x}_{2N} = f_N) \\ &= \frac{P(\overline{x}_{1N} \leq m_0 \wedge \overline{x}_{2N} = f_N)}{P\{\overline{x}_{2N} = f_N\}} \end{aligned} \qquad (4)$$

First we look at the relatively simple denominator of (4). Applying Theorem 1 yields

$$\begin{aligned} P\{\overline{x}_{2N} = f_N\} &= P\{\overline{x}_{2N} \leq (n+1/2)/N\} - P\{\overline{x}_{2N} \leq (n-1/2)/N\} \\ &= \Phi(\frac{1/2}{Nh_2}) - \Phi(\frac{-1/2}{Nh_2}) + \omega_{2N}(\frac{1/2}{Nh_2}) - \omega_{2N}(\frac{-1/2}{Nh_2}) + o(\frac{1}{\sqrt{N}}) \end{aligned} \qquad (5)$$

where the error $\omega_{2N}(x)$ of the standard normal approximation at a midpoint x is given by (2). Since

$$(1-x^2)\{\varphi(x) - \varphi(-x)\} = o(1)$$

as $x \to 0$, the total error in (5) is $o(1/\sqrt{N})$ and hence, using the first-order Taylor expansion $\Phi(x) = \Phi(0) + x\varphi(0) + O(x^2)$,

$$P\{\overline{x}_{2N} = f_N\} = \frac{1}{Nh_2}\varphi(0) + o(\frac{1}{\sqrt{N}}) \qquad (6)$$

Likewise, using Theorem 1 we get for the numerator of (4)

$$P(\overline{x}_{1N} \leq m_0 \wedge \overline{x}_{2N} = f_N) = \int_{-\infty}^{u_0h_1} \int_{-1/2N}^{1/2N} \psi(x;\overline{\Sigma}/N)dx + o(\frac{1}{\sqrt{N}}) \qquad (7)$$

Next, transform x according to $z = Tx$ with

$$T = \begin{pmatrix} 1 & -\beta \\ 0 & 1 \end{pmatrix},$$

where $\beta = \sigma_{12}/\sigma_{22} = \overline{Y}/f$. Since $|T| = 1$ and $T\overline{\Sigma}T'/N$=diag(h_1^2, h_2^2), (7) can be rewritten as

$$\begin{aligned}
&\int_{-\infty}^{u_0h_1-\beta z_2}\int_{-1/2N}^{1/2N}\psi(z;T\overline{\Sigma}T'/N)dz + o(\frac{1}{\sqrt{N}}) \\
= &\int_{-1/2N}^{1/2N}\left\{\int_{-\infty}^{u_0h_1-\beta z_2}\frac{\varphi(\frac{z_1}{h_1})\varphi(\frac{z_2}{h_2})}{h_1h_2}dz_1\right\}dz_2 + o(\frac{1}{\sqrt{N}}) \\
&= \frac{1}{N}\Phi\left(\frac{u_0h_1-0}{h_1}\right)\frac{\varphi(0)}{h_2} + o(\frac{1}{\sqrt{N}}) \qquad (8)
\end{aligned}$$

Dividing (8) by (6) yields $P_0 \sim \Phi(u_0)$ as $N \to \infty$. This concludes the proof.

The following corollary is a slightly different version of Theorem 3.

Corollary 1. Let u_1 and u_2 be two arbitrary constants ($u_1 < u_2$). Define $P_{12} = P(\overline{Y}_N + u_1h_1 \le \overline{y}_s \le \overline{Y}_N + u_2h_1)$. Then under the Lindeberg condition it holds that as $N \to \infty$

$$P_{12} \sim \Phi(u_2) - \Phi(u_1)$$

Comment. Although this corollary follows from Theorem 3, it should be noted that a direct proof is as follows. By (6) and (7),

$$P_{12} = \frac{\int_{u_1h_1}^{u_2h_1}\int_{-1/2N}^{1/2N}\psi(x;\overline{\Sigma}/N)dx + o(1/\sqrt{N})}{\varphi(0)/Nh_2 + o(1/\sqrt{N})}$$

Because in the domain of integration $x_1 = O(1/\sqrt{N})$ and $x_2 = O(1/N)$, x_2 can be set equal to zero without affecting the order of the error in the numerator. In fact, the change of the integrand thus introduced is of order $\sqrt{N}$; note that $\left|\overline{\Sigma}/N\right|^{-1/2} = O(N)$. Hence, the corresponding change of the numerator is of order $1/N$. Setting $x_2 = 0$ in the integrand and using $N\sigma^{11} = 1/h_1^2$ and $\left|\overline{\Sigma}/N\right|^{1/2} = h_1h_2$, we obtain

$$\begin{aligned}
P_{12} &\sim \frac{\frac{1}{N}\int_{u_1h_1}^{u_2h_1}\exp(-x_1^2/2h_1^2)dx_1}{2\pi h_1h_2\varphi(0)/Nh_2} \\
&= \Phi(u_2) - \Phi(u_1)
\end{aligned}$$

This concludes the proof.

3. Asymptotic distributions in rejective Poisson sampling

Consider now Hájek's rejective Poisson sampling design. His estimator for the population mean, say $\widehat{\overline{Y}}_{Haj}$, is defined as $\widehat{\overline{Y}}_{PO}$ given the condition that $n_s = n$ where

$$\widehat{\overline{Y}}_{PO} = \frac{1}{N}\sum_{k=1}^{N}a_k\frac{y_k}{\pi_k}$$

$$n_s = \sum_{k=1}^{N}a_k.$$

The π_k stand for the first order inclusion probabilities and satisfy $\sum_{k=1}^N \pi_k = n$. For further details, see Hájek (1964).

In analogy with the previous section x_k and $\overline{\Sigma}_N$ become

$$\begin{aligned} x_k &= a_k \begin{pmatrix} \frac{y_k}{\pi_k} \\ 1 \end{pmatrix} \\ \overline{\Sigma}_N &= \frac{1}{N} \sum_{k=1}^N \pi_k(1-\pi_k) \begin{pmatrix} \frac{y_k^2}{\pi_k^2} & \frac{y_k}{\pi_k} \\ \frac{y_k}{\pi_k} & 1 \end{pmatrix}. \end{aligned}$$

Suppose that the matrix $\overline{\Sigma}_N$ converges to the invertible matrix $\overline{\Sigma}$ as $N \to \infty$. Furthermore, define P_{Haj} by

$$\begin{aligned} P_{Haj} &= P(\widehat{\overline{Y}}_{Haj} \leq m_0) \\ m_0 &= \overline{Y}_N + u_0 h_1 \\ h_1^2 &= \frac{|\overline{\Sigma}|}{N\sigma_{22}} \quad (= \frac{1}{N\sigma^{11}}), \end{aligned}$$

The following theorem generalizes Theorem 3 for unequal inclusion probabilities.

Theorem 4. As $N \to \infty$ it holds under the Lindeberg condition that $P_{Haj} \sim \Phi(u_0)$. When N is sufficiently large, h_1^2 can be approximated by

$$\begin{aligned} h_{1N}^2 &= \frac{1}{N^2} \sum_{k=1}^N \pi_k(1-\pi_k)(\frac{y_k}{\pi_k} - \mu_{y^*})^2 \\ \mu_{y^*} &= \frac{\sum_{k=1}^N (1-\pi_k) y_k}{\sum_{k=1}^N \pi_k(1-\pi_k)}. \end{aligned}$$

Proof. The normality part of the proof runs along the same lines as that of Theorem 3

$$\begin{aligned} P_{Haj} &= P(\overline{x}_{1N} \leq m_0 \,|\overline{x}_{2N} = f_N) \\ &= \frac{P\{\overline{x}_{1N} \leq m_0 \wedge \overline{x}_{2N} = f_N\}}{P(N\overline{x}_{2N} = n)} \\ &\sim \Phi\left(\frac{m_0 - \overline{Y}_N}{h_1}\right) = \Phi\left(u_0\right) \end{aligned}$$

In order to show that h_1^2 can be approximated by h_{1N}^2 given in the theorem, note that

$$h_1^2 = \frac{|\overline{\Sigma}|}{N\sigma_{22}} \sim \frac{|\overline{\Sigma}_N|}{N\sigma_{N,22}} \tag{9}$$

Define

$$\begin{aligned} \gamma &= \sum_{k=1}^N \pi_k(1-\pi_k) \\ \alpha_k &= \pi_k(1-\pi_k)/\gamma. \end{aligned}$$

Furthermore, we have

$$\begin{aligned}|\overline{\Sigma}_N| &= \frac{\gamma^2}{N^2}\{\Sigma\alpha_k\frac{y_k^2}{\pi_k^2} - (\Sigma\alpha_k\frac{y_k}{\pi_k})^2\} \\ &= \frac{\gamma^2}{N^2}\sum_{k=1}^{N}\alpha_k\left(\frac{y_k}{\pi_k} - \Sigma\alpha_k\frac{y_k}{\pi_k}\right)^2 \qquad (10) \\ N\sigma_{N,22} &= \gamma, \qquad (11)\end{aligned}$$

Now it follows from (9)-(11) that

$$h_1^2 \sim \frac{|\overline{\Sigma}_N|}{N\sigma_{N,22}} = \frac{1}{N^2}\sum_{k=1}^{N}\pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k} - \mu_{y^*}\right)^2 = h_{1N}^2.$$

This concludes the proof.

It is noteworthy that approximation h_{1N}^2 in Theorem 3 is equivalent to variance approximation (8.10) of Hájek (1964). His derivation is based on an approximation of the second order inclusion probabilities π_{kl} leading to some tedious algebra. Also note that the actual inclusion probabilities, say π_k^*, for conditional Poisson sampling need not be equal to the π_k corresponding to the unconditional Poisson sampling design. As pointed out by Hájek (1964), only asymptotically it holds that $\pi_k^*/\pi_k \to 1$ as $\gamma \to \infty$. This is a maybe somewhat counterintuitive difference with equal probability sampling described in Section 2 where $\pi_k^* = \pi_k = f_N$ (k=1,...,N).

References

Basu, A.K. (2004). *Measure Theory and Probability.* Prentice Hall of India, Delhi.

Bellhouse, D.R. (2001). The central limit theorem under simple random sampling. *The American Statistician*, 55, 352−357.

Erdös, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei,* Trudy Publications, 4, 49−57.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. II. Wiley and Sons, New York.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei,* Trudy Publications, 5, 361−374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491−1523.

Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535−545.