Statistics & Probability Letters 4 (1986) 39–41
North-Holland

January 1986

A STATISTIC FOR MEASURING THE BALANCE OF A SAMPLE

Richard W. ANDREWS

The University of Michigan, Ann Arbor, MI 48109, USA

Received May 1985

Abstract: Consider a finite population which has many auxiliary variables. A statistic, which is a function of the moments of the auxiliary variables, is proposed to measure the balance of a sample. The mean and variance of this statistic are derived.

Keywords: balanced samples, finite population, auxiliary variables.

The population consists of N units, labelled $1, 2, \ldots, N$. Associated with each unit there are P variables, $x_1, x_2, \ldots, x_P$. The x's are the auxiliary variables and their values are known on all N units. Also associated with each unit are numerous variables with unknown values. These variables will be referred to as target variables and the symbol Y will be used to denote any one of them. A sample s, of n units, will be selected. The values of Y for the n sample units will be available for observation. The n units sampled will constitute a database. Requests for data on Y will be made. The nature of the requests are unknown at the time of sampling. The resulting Y data from the n units will be used to make inference about the $N-n$ unobserved units.

The three purposes of this note are (i) to define a statistic which measures the balance of a sample; (ii) to find the mean and variance of that statistic, and (iii) to recommend that this statistic along with its mean and variance be reported for multipurpose samples.

The literature in finite population sampling results for balanced and approximately balanced samples (e.g. Royall and Herson (1973), Royall and Cumberland (1981), Royall and Pfeffermann (1982) and Tam and Chan (1984)). Various measures of balance are suggested in these reports but the idea remains the same: a sample should be representative of the known variables. The cited research reports ask the question: "How does a balanced sample affect the properties of estimators?" The balancing statistic defined in this note is under investigation as to its effects on the properties of estimators. At present, the statistic should be considered a descriptive statistic for which the mean and variance are known.

The assumed sampling environment has three characteristics: (1) The sample size is small relative to the population size, (2) the sample will be used for many purposes which are not known at the time of sampling, and (3) there are many auxiliary variables.

The population univariate moments about the origin for the auxiliary variables are:

$$\mu_r(p) = N^{-1}\sum_N x_p^r,$$

for $r = 1, 2, \ldots, R$ and $p = 1, 2, \ldots, P$. The Σ_N indicates a summation over all units in the population. The population bivariate moments about the origin are:

$$\mu_{r,s}(p, q) = N^{-1}\sum_N x_p^r x_q^s,$$

This research was supported by the Air Force Office of Scientific Research and the Southeastern Center for Electrical Engineering Education under contract F49620-82-C-0035.

for $r, s = 1, 2, \ldots, R$ and $p, q = 1, 2, \ldots, P$. Since the values of the auxiliary variables are known, these population moments can be calculated.

The univariate sample moments about the origin are:

$$m_r(p) = n^{-1} \sum_s x_p^r,$$

for $r = 1, 2, \ldots, R$ and $p = 1, 2, \ldots, P$. The Σ_s indicates a summation over sampled units.

Under simple random sampling,

$$E[m_r(p)] = \mu_r(p).$$

If N is large relative to n, the population size can be considered infinite and the finite population correction can be ignored. From Kendall and Stuart (1963, p. 229),

$$\operatorname{Var}[m_r(p)] = n^{-1}[\mu_{2r}(p) - \mu_r^2(p)]. \tag{1}$$

Other expected value results are:

$$E[m_r(p)m_s(q)] = n^{-1}[\mu_{r,s}(p, q) + (n-1)\mu_r(p)\mu_s(q)]. \tag{2}$$

$$E[m_r^2(p)m_s(q)] = n^{-2}[\mu_{2r,s}(p, q) + (n-1)\mu_{2r}(p)\mu_s(q) + 2(n-1)\mu_{r,s}(p, q)\mu_r(p) + (n-1)(n-2)\mu_1^2(p)\mu_s(q)]. \tag{3}$$

$$\begin{aligned} E[m_r^2(p)m_s^2(q)] = n^{-3}[&\mu_{2r,2s}(p, q) + (n-1)\mu_{2r}(p)\mu_{2s}(q) + 2(n-1)\mu_{2r,s}(p, q)\mu_s(q) \\ &+ (n-1)(n-2)\mu_{2r}(p)\mu_s^2(q) + 2(n-1)\mu_r(p)\mu_{r,2s}(p, q) \\ &+ (n-1)(n-2)\mu_r^2(p)\mu_{2s}(q) + 2(n-1)\mu_{r,s}^2(p, q) \\ &+ 4(n-1)(n-2)\mu_r(p)\mu_{r,s}(p, q)\mu_s(q) \\ &+ (n-1)(n-2)(n-3)\mu_r^2(p)\mu_s^2(q)]. \end{aligned} \tag{4}$$

As a measure of balance define:

$$B = \sum_{r=1}^{R} \sum_{p=1}^{P} \frac{n[m_r(p) - \mu_r(p)]^2}{\mu_{2r}(p) - \mu_r^2(p)}.$$

The quantity B measures the standardized squared error between the sample and population moments for P auxiliary variables and R moments. The purpose of the remaining part of this note is to derive the mean and variance of B.

The result is:

$$E[B] = RP \tag{5}$$

and

$$\operatorname{Var}[B] = \sum_{r=1}^{R} \sum_{s=1}^{R} \sum_{p=1}^{P} \sum_{q=1}^{P} [\mu_{2r}(p) - \mu_r^2(p)]^{-1}[\mu_{2s}(q) - \mu_s^2(q)]^{-1}[f_1(\mu) + n^{-1}f_2(\mu)], \tag{6}$$

for which

$$f_1(\mu) = 2[\mu_{r,s}(p, q) - \mu_r(p)\mu_s(q)]^2,$$

and

$$f_2(\mu) = \mu_{2r,2s}(p, q) - \mu_{2r}(p)\mu_{2s}(q) - 2\mu_{2r,s}(p, q)\mu_s(q) + 2\mu_{2r}(p)\mu_s^2(q) - 2\mu_r(p)\mu_{r,2s}(p, q) + 2\mu_r^2(p)\mu_{2s}(q) - 2\mu_{r,s}^2(p, q) + 8\mu_r(p)\mu_{r,s}(p, q)\mu_s(q) - 6\mu_r^2(p)\mu_s^2(q).$$

The main steps in deriving this result are as follows. For (5),

$$E[B] = \sum_{r=1}^{R}\sum_{p=1}^{P} \frac{nE[m_r(p) - \mu_r(p)]^2}{\mu_{2r}(p) - \mu_r^2(p)} = \sum_{r=1}^{R}\sum_{p=1}^{P} 1 = RP.$$

The main steps in deriving (6) are:

$$\begin{aligned}
\mathrm{Var}[B] &= \mathrm{Var}\left\{ \sum_{r=1}^{R}\sum_{p=1}^{P} \frac{n[m_r(p) - \mu_r(p)]^2}{\mu_{2r}(p) - \mu_r^2(p)} \right\} \\
&= \sum_{r=1}^{R}\sum_{s=1}^{R} \mathrm{Cov}\left\{ \sum_{p=1}^{P} \frac{n[m_r(p) - \mu_r(p)]^2}{\mu_{2r}(p) - \mu_r^2(p)}, \sum_{p=1}^{P} \frac{n[m_3(p) - \mu_s(p)]^2}{\mu_{2s}(p) - \mu_s^2(p)} \right\} \\
&= \sum_{r=1}^{R}\sum_{s=1}^{R}\sum_{p=1}^{P}\sum_{q=1}^{P} \mathrm{Cov}\left\{ \left[\frac{n[m_r(p) - \mu_r(p)]^2}{\mu_{2r}(p) - \mu_r^2(p)} \right], \left[\frac{n[m_s(q) - \mu_s(q)]^2}{\mu_{2s}(q) - \mu_s^2(q)} \right] \right\} \\
&= \sum_{r=1}^{R}\sum_{s=1}^{R}\sum_{p=1}^{P}\sum_{q=1}^{P} n^2[\mu_{2r}(p) - \mu_r^2(p)]^{-1}[\mu_{2s}(q) - \mu_s^2(q)]^{-1} \\
&\quad \times \mathrm{Cov}\{[m_r(p) - \mu_r(p)]^2, [m_s(q) - \mu_s(q)]^2\}, \qquad (7)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Cov}\{[m_r(p) - \mu_r(p)]^2, [m_s(q) - \mu_s(q)]^2\} \\
&\quad = E\{[m_r(p) - \mu_r(p)]^2[m_s(q) - \mu_s(q)]^2\} \\
&\qquad - E[m_r(p) - \mu_r(p)]^2 E[m_s(q) - \mu_s(q)]^2.
\end{aligned}$$

By using the expected value results in (2), (3), (4), and the variance of $m_r(p)$ as given by (1), the covariance of the last equation can be found as a function of the population moments. Substituting this value for the covariance in (7) yields result (6).

For multipurpose samples which are selected in the presence of auxiliary variables it is recommended that the value of B, along with its mean, and variance be reported. This will provide information concerning the balance of that sample to potential users of the data.

References

Royall, R.M. and J. Herson (1973), Robust estimation in finite populations I., *J. Amer. Stat. Assoc.* **68**, 880–889.

Royall, R.M. and W.G. Cumberland (1981), An empirical study of the ratio estimator and estimators of its variance, *J. Amer. Stat. Assoc.* **76**, 66–88.

Royall, R.M. and D. Pfeffermann (1982), Balanced samples and robust Bayesian inference in finite population sampling, *Biometrika* **69**, 401–410.

Tam, S.M. and N.N. Chan (1984), Screening of probability samples, *Int. Statist. Rev.* **52**, 301–308.