THE EVALUATION OF DIFFERENT COUNTING RULES AND WEIGHTING PROCEDURES FOR SURVEYS WITH MULTIPLICITY

Gad Nathan, Hebrew University, Jerusalem

1. Introduction and Notation

In a sample survey with multiplicity, (Sirken [3]), designed to estimate the total number, N, of events occurring in a population of L reporting units, each event is linked to one or more reporting units, by a well defined counting rule. The counting rule is defined by the values of the indicator variable, $\delta_{\alpha,i}$, which takes the value one if the α-th event is linked to the i-th reporting unit (α=1,...,N; i=1,...,L), where:

$$(1.1) \quad s_\alpha = \sum_{i=1}^{L} \delta_{\alpha,i} \geq 1 \quad (\alpha=1,\ldots,N),$$

i.e. each event is linked to at least one reporting unit. The value of s_α is termed the multiplicity of the α-th event.

A probability sample, S, of reporting units is defined by the indicator variable $d_i(S)$, which equals 1 if the i-th reporting unit is in the sample, S.

Initially, it will be assumed that, if $d_i(S) = 1$, the vector $\underline{\delta}_\alpha = (\delta_{\alpha,1},\ldots,\delta_{\alpha,L})$ is known for all α=1,...,N, such that $\delta_{\alpha,i} = 1$, i.e., each reporting unit in the sample reports, for each event linked to it, on all other reporting units linked to the same event and there are no response errors.

If $E[d_i(S)] > 0$ (i=1,...,L), then a weighted linear multiplicity estimator of N is defined as:

$$(1.2) \quad \hat{N}(S) = \sum_{i=1}^{L} \frac{d_i(S)}{E[d_i(S)]} \sum_{\alpha=1}^{N} w_{\alpha,i}(S)\delta_{\alpha,i} ,$$

where $w_{\alpha,i}(S)$ are any real numbers. The expectation of the estimator is:

$$(1.3) \quad E[\hat{N}(S)] = \sum_{i=1}^{L} \sum_{\alpha=1}^{N} E[w_{\alpha,i}(S) \mid d_i(S) = 1]\delta_{\alpha,i} = \sum_{\alpha=1}^{N} \sum_{i=1}^{L} w_{\alpha,i}\delta_{\alpha,i} ,$$

where $w_{\alpha,i} = E[w_{\alpha,i}(S) \mid d_i(S) = 1]$.

Thus, in order that $\hat{N}(S)$ be an unbiased estimator of N for any matrix, $\Delta = ||\delta_{\alpha,i}||$ (α=1,...,N;i=1,...,L), for which (1.1) holds, the following relationship must hold;

$$(1.4) \quad \sum_{i=1}^{L} w_{\alpha,i}\delta_{\alpha,i} = 1 \quad (\alpha=1,\ldots,N).$$

Sirken [3] has proposed using the weights:

$$(1.5) \quad w_{\alpha,i}(S) = w_{\alpha,i} = 1/s_\alpha ,$$

for all S, and for all (α,i) such that $\delta_{\alpha,i} = 1$ (i.e. weighting by the reciprocal of the multiplicity). For these weights, referred to in [4] as unit weights, (1.4) obviously holds, so that N(S) is an unbiased estimator. In the following, it will be shown that, in a certain sense, this weighting is optimal for simple random sampling under the above assumption, in the absence of response errors. However alternative weighting procedures must be considered if there are response errors. The components of mean square error, taking response errors into account, are evaluated approximately as a function of certain parameters. This allows the comparison of alternative counting rules and weighting methods.

2. Optimal Weighting in the Absence of Response Errors

In the following, simple random sampling without replacement, of size ℓ, will be assumed, i.e.:

$$(2.1) \quad E[d_i(S)] = \ell/L;$$

$$E[d_i(S)d_j(S)] = [\ell(\ell-1)]/[L(L-1)], \ (i\neq j).$$

Let:

$$(2.2) \quad X_i(S) = \sum_{\alpha=1}^{N} w_{\alpha,i}(S)\delta_{\alpha,i}$$

and

$$(2.3) \quad X_i = E[X_i(S) \mid d_i(S)=1] = \sum_{\alpha=1}^{N} w_{\alpha,i}\delta_{\alpha,i} .$$

Then it can easily be shown that the variance of the estimate (1.3) is:

$$(2.4) \quad Var[\hat{N}(S)] = (L/\ell)^2 E\{\sum_{i=1}^{L} d_i(S)[X_i(S)-X_i]\}^2 + \frac{L(L-\ell)}{\ell(L-1)} \sum_{i=1}^{L} (X_i - N/L)^2$$

Each of the two terms is non-negative and, for given values of X_i, the variance is minimal if, for all S such that $d_i(S) = 1$,

$$(2.5) \quad X_i(S)-X_i = \sum_{\alpha=1}^{N} [w_{\alpha,i}(S)-w_{\alpha,i}]\delta_{\alpha,i} = 0.$$

But, for (2.5) to hold it is sufficient if, for all S, such that $d_i(S) = 1$:

$$(2.6) \quad w_{\alpha,i}(S) = w_{\alpha,i} \quad (\alpha=1,\ldots,N;i=1,\ldots,L).$$

Since, when (2.6) holds, the variance (2.4) depends only on the values of X_i, which in turn depend only on the values of $w_{\alpha,i}$, weighting which is independent of the sample is optimal, in the above sense.

For sample-independent weighting (i.e. if (2.6) holds) the variance (2.4) is minimized if $\sum_{i=1}^{L} X_i^2$ is minimal subject to (1.4). For given Δ, weights $w_{\alpha,i}$ can easily be found to attain this. However Δ must be assumed unknown, except for the row vectors $\underline{\delta}_\alpha=(\delta_{\alpha,1},\ldots,\delta_{\alpha,L})$, such that

$\sum_{i=1}^{L} d_i(S)\delta_{\alpha,i} \geq 1$ (i.e. for events linked to at least one reporting unit in the sample). But sample-independent weighting requires that for given α_o, the weights, $w_{\alpha_o,i}$, must be a function only of $\underline{\delta}_{\alpha_o}$. Thus the following minimax approach is implied:

Let $D(\underline{\delta}_{\alpha_o})$ be the set of matrices, Δ, for which the α_o-th row is $\underline{\delta}_{\alpha_o}$. For a given matrix Δ, let $W(\Delta)$ be the set of all vector functions:

$$\underline{w}: \{0,1\}^L \to R^L,$$

for which (1.4) holds, i.e.:

$$(2.7) \quad \sum_{i=1}^{L} w_i(\underline{\delta}_\alpha)\delta_{\alpha,i} = 1.$$

Let:

$$(2.8) \quad f(\underline{w},\Delta) = \sum_{i=1}^{L} \left[\sum_{\alpha=1}^{N} w_i(\underline{\delta}_\alpha)\delta_{\alpha,i}\right]^2.$$

Then, for a given vector $\underline{\delta}_{\alpha_o}$, the optimal weight vector, $\underline{w}^*(\underline{\delta}_{\alpha_o})$, is defined as that for which:

$$(2.9) \quad \min_{\underline{w}\in W(\Delta)} \max_{\Delta\in D(\underline{\delta}_{\alpha_o})} f(\underline{w},\Delta),$$

is attained.

It is easy to see that, for a given $\underline{\delta}_{\alpha_o}$,:

$$(2.10) \quad \max_{\Delta\in D(\underline{\delta}_{\alpha_o})} f(\underline{w},\Delta) = (N-1)^2 + \sum_{i=1}^{L} w_i^2(\underline{\delta}_{\alpha_o}) + 2(N-1) \max_{1\leq i\leq L} w_i(\underline{\delta}_{\alpha_o}),$$

for any $\underline{w}\in W(\Delta)$ such that $\Delta\in D(\underline{\delta}_{\alpha_o})$. The maximum is attained by Δ^* such that $\underline{\delta}^*_{\alpha_o} = \underline{\delta}_{\alpha_o}$ and, if

$$\max_{1\leq i\leq L} w_i(\underline{\delta}_{\alpha_o}) = w_{i_o}(\underline{\delta}_{\alpha_o}),$$

$$(2.11) \quad \delta^*_{\alpha,i} = \begin{cases} 1 ; & i=i_o \\ 0 ; & i\neq i_o \end{cases},$$

$$(\alpha\neq\alpha_o;\ \alpha=1,\ldots,N).$$

The minimum of (3.4), for $\underline{w}\in W(\Delta^*)$, is:

$$(2.12) \quad \min_{\underline{w}\in W(\Delta^*)} f(\underline{w},\Delta^*) = (N-1)^2 + \frac{1}{s_{\alpha_o}} + \frac{2(N-1)}{s_{\alpha_o}}.$$

This minimum is attained by $\underline{w}^*$ for which $w_i^*(\underline{\delta}_{\alpha_o}) = \frac{1}{s_{\alpha_o}}$, for all i=1,...,L, such that $\delta_{\alpha_o i} = 1$.

Thus the unit weighting (1.5) is optimal in the above minimax sense.

3. Weighting in the Presence of Response Errors

If, as usually happens in practice, events, for which $\delta_{\alpha,i} = 1$ and for which $d_i(S) = 1$, are sometimes under-reported, or if there is over-reporting of events, the optimality of the weights (1.5) is not necessarily attained. For the case where the reciprocals of the multiplicities are used as weights, an approximation to the total mean square error, including response and sampling error components, of the estimate (1.3) is given in Nathan [1], under certain simplifying assumptions. In the following a similar development for a general weighting scheme will be given.

The weighting can be defined by means of a variable, $Z_{\alpha,i}$, measurable for all i, such that $\delta_{\alpha,i}=1$ for any event, α, for which $\sum_{i=1}^{L} d_i(S)\delta_{\alpha,i} \geq 1$ (i.e. the event is linked to at least one reporting unit in the sample).

Let $Z_\alpha = \sum_{i=1}^{L} Z_{\alpha,i}\delta_{\alpha,i}$. Then for the weights:

$$(3.1) \quad w_{\alpha,i} = \frac{Z_{\alpha,i}}{Z_\alpha},$$

the relationship (1.4) holds, so that, in the absence of response errors, $\hat{N}(S)$, would be unbiased. Thus if the variable $Z_{\alpha,i}$ is the number of persons linked to the event in the household, element weighting, as defined by Sirken and Royston [4], is obtained.

Response errors may occur both due to under-reporting or over-reporting of events and to errors in reporting the values of $Z_{\alpha,i}$ and of Z_α. For the present it is assumed that there is no over-reporting of events. This assumption is made to simplify the expressions and will be later relaxed.

In addition to the assumptions and notations of [1], it will be assumed that the weighting variables, $Z_{\alpha,i}$, are observed without error, if $d_i(S)=1$, i.e. the reporting unit itself reports correctly on its own value. However, response errors may occur in the reporting of Z_α. Let $Z_\alpha(i,t)$ be the value reported for Z_α by the i-th reporting unit at trial t. Unbiasedness will be assumed, so that:

$$(3.2) \quad E_t[Z_\alpha(i,t)] = Z_\alpha \quad (i=1,\ldots,L;\ \alpha=1,\ldots,N)$$

The relative response variance of $Z_\alpha(i,t)$ is assumed to be independent of α and of i, so that:

$$(3.3) \quad \mathrm{Var}_t[Z_\alpha(i,t)]/Z_\alpha^2 = V^2$$
$$(i=1,\ldots,L\ \alpha=1,\ldots,N).$$

The sample estimate of N, at trial t, for sample S, is:

$$(3.4) \quad \hat{N}(S,t) = \frac{L}{\ell} \sum_{i=1}^{L} d_i(S) \sum_{\alpha=1}^{N} w_{\alpha,i}(t)\delta_{\alpha,i},$$

where:

$$(3.5) \quad w_{\alpha,i}(t) = \frac{Z_{\alpha,i}}{Z_\alpha(i,t)}.$$

If we denote the average of the weights, $w_{\alpha,i}$, for reports of relationship r, by:

$$(3.6) \quad A_r = \frac{1}{N} \sum_{\alpha=1}^{N} \sum_{i=1}^{L} w_{\alpha,i}\delta_{\alpha,i,r}, \quad (r\varepsilon C)$$

so that $\sum_{r\varepsilon C} A_r = 1$, then, similarly to the development in (1) the bias of the estimate can be approximated by:

$$(3.7) \quad B = N_o - N \doteq -N \sum_{r\varepsilon C} \{1-(1+V^2)P_r\}A_r.$$

Neglecting the correlated response variance, the (simple) response variance, RV, can be approximated by:

$$(3.8) \quad RV \doteq \frac{L}{\ell} N \sum_{r\varepsilon C} [V^2 P_r B_r + (1+V^2)P_r(1-P_r)A_r]$$

where:

$$(3.9) \quad B_r = \frac{1}{N} \sum_{\alpha=1}^{N} \sum_{i=1}^{L} w_{\alpha,i}^2 \delta_{\alpha,i,r}$$

is the average of the squared weights, $w_{\alpha,i}^2$, for reports of relationship r. Similarly the sampling variance can be approximated by:

$$(3.10) \quad SV = \frac{L-\ell}{\ell} [N(1+V^2)^2\{ \sum_{r\varepsilon C} P_r^2 B_r$$

$$+ \sum_{r,r'\varepsilon C} \sum_{\alpha,\alpha'=1}^{N} \sum_{i=1}^{L} w_{\alpha,i}w_{\alpha',i}\delta_{\alpha,i,r}\delta_{\alpha',i,r'}\}$$

$$- N_o^2/L].$$

The above expressions for the components of the mean square error, (3.7), (3.8) and (3.10), are exactly the same, for the case of no over-reporting, as the expressions (3.2), (3.8) and (3.9) in [1], with $w_{\alpha,i}$ replacing $1/s_\alpha$. The above expressions are, in fact, a generalization of the case of unit weighting (i.e. $w_{\alpha,i}=1/s_\alpha$). They can thus be easily extended to the case where M additional "non-events" are liable to be reported in exactly the same way as in [1], with $1/s_\alpha$ replaced by $w_{\alpha,i}$.

4. Evaluation of the Components

The values of the parameters required to evaluate the components can be estimated from an evaluation survey, as specified in [1], or as proposed by Sirken and Royston [4]. However, the values of A_r and of B_r which are, of prime importance in the components of error and contribute importantly to the differences between counting rules and weighting methods, can, in some cases be obtained from approximations to the distributions of components of the weights $w_{\alpha,i}$ by simple single-parameter distributions.

For example, the empirical results of a multiplicity study on births in Israel, described fully in Nathan, Schmelz and Kenvin [2], show that certain household multiplicities were distributed approximately as Poisson distributions.

Thus, for the counting rule for which births are reported by the woman giving birth, by her sisters and by her mother (in that order of precedence, i.e. mothers only report in the absence of sisters in the household), the distributions of the number of households of sisters for networks with and without mothers (not residing with daughters) are given in Table 1. The fit to the theoretical Poisson distributions is good and the difference between the empirical and the theoretical distributions is not significant. The values of the components of relative error resulting from the use of this approximation, for various combinations of the parameters are given in Table 2, for unit weighting. A similar approximation for element weighting gives the results of Table 3.

References

[1] Nathan, Gad, "An Empirical Study of Response and Sampling Error for Multiplicity Estimates with Different Counting Rules," Journal of the American Statistical Association, Vol. 71 (1976), No. 357 (to appear).

[2] Nathan, Gad, Usiel O. Schmelz, and Jay Kenvin, "Multiplicity Study of Marriages and Births in Israel," National Center for Health Statistics, Series 2, No. 70, Washington, D.C., 1976 (in press).

[3] Sirken, Monroe, "Household Surveys with Multiplicity, Journal of the American Statistical Association," Vol. 65 (1970), pp. 257-266.

[4] Sirken, Monroe and Patricia N. Royston, "Design Effects in Retrospective Mortality Surveys." Proceedings Social Statistics Section, American Statistical Association, 1976.

Table 1: Empirical and theoretical distributions of household multiplicities (reports on births by women giving birth*)

Number of households of sisters	Separate household of mother			
	$s_{\alpha,2} = 0$ (1)		$s_{\alpha,2} = 1$ (2)	
$s_{\alpha,3}$	Empirical	Theoretical	Empirical	Theoretical
Total	164	164	164	164
0	41	37.04	52	47.85
1	54	55.11	55	58.95
2	38	41.00	34	36.30
3	18	20.33	16	14.90
4+	13	10.52	7	6.00
Mean	1.488		1.232	
χ^2 (goodness of fit)	1.517		1.017	

*Source: Multiplicity study of marriages and births in Israel.

(1) Mother's mother resides in household with daughter or deceased.

(2) Mother's mother resides in separate household (without daughters).

Table 2: Components of mean square error for various values of basic model parameters - unit weighting.

Parameter/Component	Multiplicity rule alternatives							Conventional rule
Values of basic model parameters								
Mean number of sisters' households								
per network without household of mother - λ_o	1.50	1.80	1.50	1.80	1.50	1.50	1.50	---
per network with household of mother - λ_1	1.25	1.25	1.50	1.50	1.25	1.25	1.25	---
Proportion of networks without household of mother - R	.50	.50	.50	.50	.60	.50	.50	---
Under-reporting probabilities - $1-P_r$								
Event household (r=1)	.02	.02	.02	.02	.02	.02	.02	.02
Household of mother (r=2)	.09	.09	.09	.09	.09	.09	.09	---
Households of sisters (r=3)	.19	.19	.19	.19	.19	.15	.19	---
Over-reporting rates - $Q_r^{(1)}$								
Event household (r=1)	.02	.02	.02	.02	.02	.02	.02	.02
Household of mother (r=2)	.18	.18	.18	.18	.18	.18	.18	---
Households of sisters (r=3)	.20	.20	.20	.20	.20	.20	.15	---
Components of relative standard error (percentages)								
Total root mean square error - $\sqrt{MSE}/N$	4.12	4.02	4.06	3.96	4.10	5.00	3.93	5.37
Bias - B/N	1.14	1.06	1.00	0.92	0.83	3.07	0.49	0.04
Response standard error - $\sqrt{RV}/N$	2.39	2.44	2.41	2.46	2.37	2.34	2.29	0.96
Sampling standard error - $\sqrt{SV}/N$	3.16	3.02	3.10	2.96	3.24	3.19	3.15	5.29

Table 3: Components of mean square error for various values of basic model parameters - element weighting.

Parameter/ Component								
	Values of basic model parameters							
Mean number of reporting persons:								
in event household and in households of sisters (excluding woman giving birth) in networks without mother	3.40	4.00	3.40	4.00	3.40	3.40	3.40	3.40
in households of sisters in networks with mother	1.25	1.25	1.50	1.50	1.25	1.25	1.25	1.25
Proportion of networks without household of mother	.50	.50	.50	.50	.60	.50	.50	.50
Proportion of sisters in separate households	.90	.90	.90	.90	.90	.95	.90	.90
Under-reporting probabilities								
Event household	.02	.02	.02	.02	.02	.02	.02	.02
Household of mother	.09	.09	.09	.09	.09	.09	.09	.09
Households of sisters	.19	.19	.19	.19	.19	.19	.15	.19
Over-reporting rates								
Event household	.02	.02	.02	.02	.02	.02	.02	.02
Household of mother	.18	.18	.18	.18	.18	.18	.18	.18
Households of sisters	.20	.20	.20	.20	.20	.20	.20	.15
	Components of relative standard error (percentages)							
Total root mean square error - $\sqrt{MSE}/N$	4.17	4.09	4.05	3.65	4.02	4.12	5.63	3.63
Bias - B/N	1.91	1.86	1.72	0.72	1.56	1.86	4.26	0.07
Response standard error - $\sqrt{RV}/N$	2.56	2.59	2.58	2.58	2.57	2.56	2.50	2.45
Sampling standard error - $\sqrt{SV}/N$	2.68	2.57	2.61	2.48	2.67	2.62	2.71	2.68