Sampling with probabilities proportional to the variable of interest

J.J.A. Moors/R.Sm eets / F.W.M. Boekema

Abstract

To estimatet he meansoj urn time, a sample of Tilburg fair visitorsw as asked for the duration of their stay on the fair grounds. The longer a visitor'soj urn, the larger his/her probability of being in terviewed will be; therefore, longer sojourn times will be overrepresented in the sample. As a consequence, the arithmetic sample mean is not a suitable estimator.

The paper places this problem against a theoretical background. As a better estimator the harmonic meanof t he observed sojourn times is presented. In addition, a variance estimator is given. The properties of these estimators are difficult to derivea nalytically.Ins ead, their beha viour is studied in a number of examples.

Keywords: harmonic mean, *pps*-sampling, *ppy*-sampling, renewal theory, sojourn time, variance estimator.

1 Introduction

The Tilburg fair is the Netherlands' largest; its economic impact, therefore, is great. A first attempt to investigate these economic features was made in SMEETS (1988). Since expenditures of visitors will tend to increase with sojourn times, an important economic indicator will be the average sojourn time of Tilburg fair visitors.

To estimate this crucial quantity, pollsters questioned nearly 2000 fairgrounds visitors in 1988 about the duration of their stay. Respondents were selected without exact instructions: pollsters moved on the fair grounds and interviewed 'randomly' selected visitors. Since the probability that a visitor is being interviewed will increase with the duration of stay, as in table visitor selection model is probability proportional to sojourn time. In other words, selection probability is proportional to the variable that is being investigated. This sampling scheme is called *ppy*-sampling; it is studied here in a more general context.

The paper shows that to estimate the average duration of stay under a *ppy*-sampling scheme, the harmonic mean of the observed sojourn times should be used rather than the arithmetic mean. An admittedly, crude - estimator for the variance is presented; it is an extremely simple function of both arithmetic and harmonic mean.

In more detail, the organisation of the paper is the following. Section 2 treats sampling with probabilities proportional to size, where 'size' is a fully known variable. To avoid unnecessarily laborious notations, it will be assumed in the beginning that all random variables are absolutely continuous. Adaptation of the notation to the discrete case is straightforward, as some of the examples illustrate.

In Section 3, *ppy*-sampling is defined and studied; of course, now the values y_i of the variable of interest, are assumed to be unknown. The properties of the proposed estimator are discussed, mostly by means of three examples. Section 4 considers variance estimation, while Section 5 applies the theory developed to the Tilburg fair problem, that triggered this research. The final Section 6 briefly presents comparable outcomes for the Tilburg 1995 fair (VERMEULEN, 1995); of course, the methodology is applicable to more general gatherings of large crowds. The relation with renewal theory is pointed out.

2 Sampling with probabilities proportional to size

Consider a pair of absolutely continuous random variables (X,Y) with joint density $f : \mathbb{R}^2 \to \mathbb{R}$. The marginal distributions are denoted by f_X and f_Y. For any function $k : \mathbb{R}^2 \to \mathbb{R}$, we define

$$\mu_k = E[k(X,Y)] = \iint k(x,y) f(x,y)\, dx dy$$

provided this expectation exists; e.g.

$$\mu_x = \iint x f(x,y)\, dx dy = \int x f_X(x) dx.$$

Any random observation of the pair (X,Y) has f as density. Consider the case, however, that observations are not made at random, but with probabilities proportional to a known positive function h of Y with existing mean μ_h. Then the function p defined by

$$p(y) = h(y)/\mu_h \tag{1}$$

reweights the density f. An observation of (X,Y) now leads to a new pair (V,W) of random variables taking the same values as (X,Y), but with a density g given by

$$g(x,y) = f(x,y)p(y) \tag{2}$$

The marginal densities are

$$g_W(y) = f_Y(y)p(y)$$

$$g_V(x) = \int f(x,y)p(y)dy$$

Of course, $E[k(V,W)] \neq E[k(X,Y)] = \mu_k$ - unless p is constant (the case of random sampling). However,

$$E[k(V,W)/p(W)] = \mu_k \tag{3}$$

So, first dividing the observation $k(V,W)$ by the sampling probability and then summing gives an unbiased estimator for μ_k. This fact is well-known from sampling with unequal probabilities in finite population sampling; the complete distribution of the auxiliary variable Y is assumed known. See for example, HEDAYAT & SINHA (1991) or SÄRNDAL et al. (1992). If h is the identity, this sampling scheme is called *pps*-sampling: sampling proportional to size (Y); Y is assumed to be positive.

In this important special case $h(y)=y$, (1) is simplified to

$$p(y)=y/\mu_y \tag{4}$$

An immediate consequence is

$$E(W)=E(Y^2)/\mu_y,\quad E(V)=E(XY)/\mu_y \tag{5}$$

while of course (2) continues to hold.

(Note that *pps*-sampling is not the only case of interest. Consider a case with balls of different radii in Y. Balls are selected by choosing a point in $\mathbb{R}^3$ with random coordinates; if this point is contained in one of the balls, this ball is sampled. In this case $p(y)\propto y^3$.)

EXAMPLE 1. Consider the following discrete bivariate distribution of (X,Y) with probability mass function f. Then marginal distributions and some parameters are added.

	$f(x,y)$			
	y 1	2	4	$f_X(x)$
x				
0	0.2	0.3	–	0.5
1	–	0.4	0.1	0.5
$f_Y(y)$	0.2	0.7	0.1	1

$$\begin{aligned}\mu_x &= 0.5\\ \sigma_x^2 &= 0.25\\ \mu_y &= 2\\ \sigma_y^2 &= 0.6\\ \sigma_{xy} &= 0.2\\ \rho_{xy} &= 0.5164\end{aligned}$$

In sampling with probability to size Y, p follows from (4):

y	1	2	4
$p(y)$	0.5	1	2

leading to the following joint distribution of observations (V,W).

$g(v,w)$				
	w	1	2	4	$g_V(v)$
v					
0		0.1	0.3	–	0.4
1		–	0.4	0.2	0.6
$g_W(w)$		0.1	0.7	0.2	1

$$\begin{aligned} \mu_v &= 0.6 \\ \sigma_v^2 &= 0.24 \\ \mu_w &= 2.3 \\ \sigma_w^2 &= 0.81 \\ \sigma_{vw} &= 0.22 \\ \rho_{vw} &= 0.4990 \end{aligned}$$

It is easy to check that (5) holds. On the other hand,

$$W/p(W) = 2 \Rightarrow E[W/p(W]=2= \quad \mu_y$$

while the distribution of $Z = V/p\,(W)$ is given by

z	0	0.5	1
$f_Z(z)$	0.4	0.2	0.4

so that $E(Z) = E[V/p\,(W)] = 0.5= \quad \mu_x$. This is in agreement with the general result (3). □

In general, Y is of interest only as an auxiliary variable determining the sampling probabilities, while X is the variable under investigation. However, in some instances, the 'size' Y itself is the variable of interest.

Further, attention will be concentrated as a rule on the original variables X and Y. It is important to note, however, that the observed variables (V, W) may be of importance too.

Example 2 illustrates both remarks.

EXAMPLE 2. In a small school with fourty pupils four classes are formed with 4, 6, 12 and 18 pupils, respectively. For this school, class-size Y has mean $\mu_y = 10$. Draw one pupil at random and ask for the number of pupils in his/her class. This gives the following functions.

y	4	6	12	18
$f_Y(y)$	0.25	0.25	0.25	0.25
$p(y)$	0.4	0.6	1.2	1.8
$f_W(y)$	0.1	0.15	0.3	0.45

For the observation W of Y it follows $E(W)=13$. This number can be interpreted as the mean number of pupils per class *as experienced by the pupils*. Note that it is the 'quadratic mean' $E(Y^2)/E(Y)$, in agreement with (5). (In slightly otherwords this outcome can be read as: for the population of fourty pupils the mean number of classmates in 12.) □

3 Sampling with probabilities proportional to the variable of interest

From now on, attention will be concentrated on Y being both the auxiliary 'size' variable and the variable under investigation. To make the problem statistically interesting, the probability distribution of Y is supposed to be completely unknown. The only information is obtained through observations W which are drawn with probabilities $p(w)= h(w)/\mu_h; \mu_h = E[h(Y)]$ is the unknown estimand.

An estimator for μ_h can be found by applying (3) to the (constant) function $k(V,W)=1$. This gives

$$1 = E[1/p(W)] = \mu_h E[1/h(W)]$$

or

$$1/E[1/h(W)] = \mu_h \tag{6}$$

This suggests the following estimator M_h for μ_h, based on n observations $W_1, W_2, \ldots, W_n$:

$$M_h = \frac{1}{(1/n)\sum_1^n 1/h(W_i)} = \frac{n}{\sum_1^n 1/h(W_i)} \tag{7}$$

i.e. the harmonic mean of the $h(W_i)$.

The special case that arises if h is the identity will be called sampling with probabilities proportional to the variable of interest Y, abbreviated *ppy*-sampling. The estimator (7) for μ_y then is the harmonic mean M of the observations W_i.

According to Basu's theorem, M_h is a consistent estimator. However, other statistical properties of M_h are hard to establish in general. Therefore, only three special cases will be considered.

EXAMPLE 3. Assume that Y has the inverse gamma distribution $I\Gamma(\lambda, \rho)$ with $\rho > 2$ implying

$$f(y) = \frac{\lambda^\rho}{\Gamma(\rho)} y^{-(\rho+1)} e^{-\lambda/y}, \quad y > 0.$$

Then $\mu_y = \lambda/(\rho - 1)$ and $\sigma_y^2 = \lambda^2/(\rho - 1)^2(\rho - 2)$. The distribution of an observation W, obtained by *ppy*-sampling - compare (2) and (4), follows:

$$W \sim I\Gamma(\lambda, \rho - 1).$$

Standard properties of gamma and inverse gamma distributions lead to the following successive results:

$$1/W \sim \Gamma(\lambda, \rho - 1)$$

$$\textstyle\sum_{i=1}^n 1/W_i \sim \Gamma[\lambda, n(\rho - 1)]$$

$$1/\textstyle\sum_{i=1}^n 1/W_i \sim I\Gamma[\lambda, n(\rho - 1)]$$

$$M \sim I\Gamma[n\lambda, n(\rho - 1)].$$

Immediate results are

$$E(M) = \frac{\lambda}{\rho - 1 - 1/n}, \quad V(M) = \frac{\lambda^2}{n(\rho - 1 - 1/n)^2(\rho - 1 - 2/n)}.$$

So for large n, the bias $B(M)$ and variance of the estimator M for μ_y are

$$B(M) \doteq \lambda/[n(\rho - 1)^2], \quad V(M) \doteq \lambda^2/[n(\rho - 1)^3]$$

approximately. □

EXAMPLE 4. Here, Y is a dichotomous variable; *ppy*-sampling leads to the following probability mass function f_W of W.

y	$f_Y(y)$	$p(y)$	$f_W(y)$	
1	2/3	3/4	1/2	$\mu_y = 4/3, \quad \sigma_y^2 = 2/9$
2	1/3	3/2	1/2	$\mu_w = 3/2, \quad \sigma_w^2 = 1/4$

Let T denote the number of observations $W_1, W_2, \ldots, W_n$ having the value 1. Then $T \sim B(n, 1/2)$ and

$$M = \frac{n}{T + (n-T)/2} = \frac{2n}{n+T}$$

Expectation and variance of M then follow immediately from

$$E(M^k) = \frac{1}{2^n} \sum_{t=0}^{n} \left(\frac{2n}{n+t}\right)^k \binom{n}{t}.$$

Table 1 shows some numerical results. The last column will be discussed in the next Section.

Table 1 Moments of the harmonic mean in Example 4.

n	$E(M)$	$nV(M)$	$E(S^2)$	n	$E(M)$	$nV(M)$	$E(S^2)$
2	1.4167	0.2639	0.1111	50	1.3363	0.2011	0.2192
3	1.3875	0.2508	0.1538	100	1.3348	0.1993	0.2207
5	1.3648	0.2323	0.1853	500	1.3336	0.1979	0.2219
10	1.3486	0.2153	0.2056	1000	1.3335	0.1977	0.2221

For large n, M is nearly unbiased for $\mu_y = \frac{4}{3}$ with variance $0.198/n$. □

EXAMPLE 5 (see EXAMPLE 2). Assume that from the population of fourty pupils n pupils are drawn at random (with replacement). Then the mean class-size $\mu_y (= 10)$ can

be estimated by the harmonic mean M of the class-sizes obtained from the sample. In this case, the behaviour of M is studied by means of simulation.

For different values of $n, k = 500$ (*ppy*-) samples were simulated; denote the outcome of M in sample j ($j = 1, 2, \ldots, k$) by m_j. Table 2 shows the quantities

$$\bar{m} = \frac{1}{k}\sum_{j=1}^{k} m_j, \quad v(M) = \frac{1}{k-1}\sum_{j=1}^{k}(m_j - \bar{m})^2. \tag{8}$$

They can be viewed as approximations to the parameters $E(M)$ and $V(M)$, respectively. The last column of the table will be discussed in the next Section.

Table 2 Simulated means $\bar{m}$ and variances $v(M)$ of the harmonic mean in Example 2.

n	$\bar{m}$	$nv(M)$	s^2	n	$\bar{m}$	$nv(M)$	s^2
2	11.789	38.322	10.246	50	10.068	41.072	29.171
3	11.374	42.337	15.716	100	10.047	38.863	29.601
5	10.845	43.825	21.230	500	10.009	39.714	29.896
10	10.369	42.792	26.003	1000	10.001	39.658	29.955

For large n, M is nearly unbiased with approximate variance $39.7/n$. □

4 Estimating the variance of M

No statistical analysis is complete without a variance estimate. So the next two steps will be: to find an approximate expression for the variance of M and to derive an estimator. Our solution starts with an alternative formulation of M.

The sample gives a picture of the distribution of W. To obtain a picture of the distribution of the original variable Y, the observations W_i must be reweighted; because of the *ppy*-sampling plan the weights

$$G_i = \frac{1/W_i}{\sum^n 1/W_i}$$

are appropriate. Mean and variance of this re-weighted sample are given by

$$\sum_{i=1}^{n} G_i W_i (= M), \quad \sum_{i=1}^{n} G_i (W_i - M)^2.$$

So, the second statistic S^2, can be used as a consistent estimator for σ_y^2. Since M is a (arithmetic) sample mean with - admittedly - rather peculiar weights, a (crude) approximation for its variance will be given by

$$V(M) \doteq \sigma_y^2/n. \tag{9}$$

Note that S^2 can be rewritten as

$$S^2 = [\sum_{i=1}^{n}(W_i - M)^2/W_i]/\sum_{i=1}^{n} 1/W_i = \frac{M}{n}[\sum_{i=1}^{n} W_i - nM].$$

Introducing $\bar{W} = \frac{1}{n}\sum^n W_i$, this estimator for σ_y^2 becomes

$$S^2 = M(\bar{W} - M). \tag{10}$$

For $n = 1000$, the following values were found in Example 4:

σ_y^2	$nV(M)$	$E(S^2)$
0.222	0.198	0.222

and in Example 5:

σ_y^2	$nv(M)$	s^2
30	39.7	30.0

where

$$s^2 = \frac{1}{k}\sum_{j=1}^{k} s_j^2, \quad s_j^2 = m_j(\bar{w}_j - m_j).$$

Indeed, (9) only offers a crude approximation, which however may be useful in practice. The search for a better variance estimator will be continued.

5 Application

The Tilburg fair is the Netherlands' largest. It takes nine days - at the end of July; the number of attractions exceeds 200. Among them are games of skill like shooting galleries, fairground attractions like Ferris wheels, lotteries and gambling halls. Apart from its entertaining and cultural features, the economic importance is rather impressive: e.g., the city of Tilburg receives over 2 mln Dutch guilders from the showmen. This amount is based in particular on the number of visitors, the average sojourn time and their total expenditures. Reliable estimates of these quantities therefore are of great importance, both for individual showmen and for the city of Tilburg.

The first attempt to obtain a detailed picture of the economic impact of the Tilburg fair was reported in SMEETS (1988). Here we concentrate on one feature of this project: estimating the mean sojourn time of the visitors of the fairgrounds.

During the fair, pollsters walked on the fairgrounds and asked nearly 2000 visitors to answer a questionnaire. One of the questions was 'How long is your average stay on the fair site per visit?' The first two columns of Table 3 show the results; g_k denotes the observed frequency.

Table 3. Duration of sojourn on fairgrounds.

stay (hrs)	g_k	w_k	$g_k w_k$	g_k^*
0– <1	104	0.5	52	208
1– <2	309	1.5	463.5	206
2– <3	556	2.5	1390	222.4
3– <4	416	3.5	1456	118.9
4– <5	265	4.5	1192.5	58.9
5– <6	121	5.5	665.5	22
6– <7	81	6.5	526.5	12.5
≥ 7	134	8	1072	16.8
Total	1986	–	6818	865.5

Source: SMEETS (1988), p. 101

From the classmids w_k in column 3 and from column 4 follows the (arithmetic) mean duration of stay in this sample:

$$\bar{w} = \sum_k g_k w_k / \sum_k g_k = 6818/1986 = 3.433.$$

If the observations had been obtained by random sampling, this value would have been a suitable estimate for μ_y, the mean sojourn time of all visitors. However, the probability of being approached by a pollster will increase with the duration of stay. Hence, a better model is obtained by assuming *ppy*-sampling.

So, a better picture of the sojourn time distribution in the population is obtained by using the reweighted frequencies $g_k^* = g_k/w_k$. This gives the estimate

$$m = \sum_k g_k^* w_k / \sum g_k^* = 1986/865.5 = 2.295.$$

which is only 67% of the arithmetic mean. The estimator (10) for the variance σ_y^2 in the population takes the value

$$s^2 = m(\bar{w} - m) = 2.612.$$

Following (9), an indication of the variance of M then is given by

$$v(m) = s^2/n = 0.00132$$

leading to the approximate 95%-confidence interval (2.224, 2.366) for μ_y.

As a check, a resampling procedure was used. From the observed distribution (i.e. columns 2 and 3 of Table 3) $k = 500$ random samples (with replacement) of size 1986 were drawn. For each sample j the harmonic mean m_j was calculated. Then (8) leads to the values

$$\bar{m} = 2.340, \quad v(M) = 0.00248$$

Note that $\bar{m}$ is within the above confidence interval; regretfully, the difference between $v(M)$ and $v(m)$ is substantial. A better variance estimator is wanted.

According to (3), the mean μ_x of any other variable X can be estimated unbiasedly from an observed pair (V, W) by means of

$$V/p(W) = \mu_y V/W.$$

Since μ_y is unknown, a natural estimator M_x for μ_x, based on n observations $(V_1, W_1), (V_2, W_2), \ldots, (V_n, W_n)$ is

$$M_x = \frac{M}{n} \sum_{i=1}^{n} V_i / W_i = \sum_{i=1}^{n} G_i V_i.$$

So, from a bivariate frequency table of (V, W), an estimate for μ_x can be obtained - compare EXAMPLE 1. Note the central role of M.

6 Discussion

To quantify the economic impact of the Tilburg fair, the average sojourn time is one of the key aspects. Together with the number of visitors it determines the fee showmen have to pay the city of Tilburg. In estimating this mean sojourn time, the size of the fair is in itself a major problem: the fair founds occupy a large area within the city center; since entrance is free, there are no clear-cut entrance/exit-gates. By consequence, it is difficult to take a random sample of visitors and this lead to the *ppy*-sample in the 1988 survey.

To avoid the problems connected with this sample design, the 1995 survey was organised at the fair's main exit roads. By questioning persons obviously leaving the fair's premises an approximately random sample of 1780 visitors was obtained.

Table 4 compares the distributions of the sojourn times in 1988 and 1995. Columns 2 and 3 are derived from Table 3 (columns 2 and 5, respectively). The last column is based on VERMEULEN (1995).

Table 4. Relative frequency distributions of sojourn times.

stay (hrs)	1988		1995
	sample	reweighted	sample
0− <1	5.2%	24.0%	17.6%
1− <3	43.6	49.5	42.1
3− <4	20.9	13.6	20.8
4− <5	13.3	6.8	7.0
5− <6	6.1	2.5	4.9
≥ 6	10.8	3.4	7.4

It is clear that the 1988 sample resembles the 1995 distribution much better after reweighting.

In recent years, many non-profit organisations are starting to operate on a more commercial basis. In particular municipalities are forced - due to budget cuts - to explore and exploit their opportunities in the fields of tourism and recreation. Quantitative data on the economic performance of these new developed mass activities are badly needed. The problem of estimating the number of participants at mass meetings is notorious: the guesses of organizers and police officials may easily differ a factor 3. The paper shows how one of the problems arising in these situations can be handled.

The subject of this paper is related to renewal theory: the first equation in (5) can be found in KOHLAS (1982), p. 57. However, the problem was considered from a sampling point of view; besides, estimators (M and S^2) were presented for mean and variance in the population. These estimators will prove to be useful in renewal situations as well.

Acknowledgment.
We are indebted to A.C.A. Mathijssen who carefully made all calculations.

References

HEDAYAT, A.S. & B.K. SINHA (1991), Design and inference in finite population sampling, Wiley New York.

KOHLAS, J. (1982), Stochastic methods of operations research, Cambridge University Press.

SMEETS, R. (1988), The economic effects of the Tilburg fair (in Dutch), master's thesis Tilburg University.

SÅRNDAL. C.-E., B. SWENSSON and J. WRETMAN (1992), Model assisted survey sampling, Springer Verlag, New York.

VERMEULEN, M. (1995), Image-survey Tilburg 1995 fair (in Dutch), Report city of Tilburg, Department of General Affairs.