CHAPTER 7

**SAMPLING WITH UNEQUAL PROBABILITIES**

**7.1. INTRODUCTION.**

Up till now equal probability sampling selection procedure and estimation methods have been discussed in previous chapters. In this Chapter those selection procedures will be considered in which probability of selection varies from unit to unit (unequal probability) in the population. In equal probability, selection does not depend how large or small that unit is but in unequal probability these considerations are made. The probabilities must be known for all units of the population.

The use and general theory of unequal probabilities in sampling was perhaps first presented by Hansen and Hurwitz (1943). They demonstrated, however, that use of unequal selection probabilities within a stratum frequently made far more efficient estimator of total than did equal probability selection provided *measure of size* **(zi** i.e. **)** is sufficiently correlated with *estimand,* variable under study Yi. A method of selection in which the units are selected with probability proportionate to given measure of size, related to the characteristic under study is called *unequal probability sampling or the probability proportional to size sampling****,***commonly known as PPS or πPS sampling.

**7.2. SAMPLING WITH UNEQUAL PROBABILITIES WITH REPLACEMENT [PPS SAMPLING].**

The technique developed by Hansen and Hurwitz (1943) is based on with-replacement selection process. They proposed a two stage sampling scheme (will be discussed in Chapter VIII). The first stage selection took place in independent draws. At each draw, a single first-stage unit is selected with probabilities proportional to a size, the number of second-stage sampling units within each first-stage units. At the second-stage, the same number of second stage-units is selected from each sampled first-stage unit. Because it is possible for the same first-stage unit to be selected more than once, this type of unequal probability sampling is generally known as sampling with replacement. Since, however, the independence of the draws is not necessary condition for the units to have a non-zero probability of being selected more than once, another name first suggested by Hartley and Rao (1962) is *multinomial sampling,* a term justified by the multinomial distribution of the number of units in the sample. Unequal probability can however be used in single stage design. This selection procedure is explained as:

A list of 523 villages along with population of males and females of Multan district is given at the end of Chapter. In order to understand the selection procedure of probability proportional to size sampling 5% sample has been selected from this population. In order to select a sample we cumulate the measure of sizes (area) under this selection procedure, 26(5% of total villages) random numbers are selected from 001 to 956204. These random numbers along with the serial number of villages, total population and initial probabilities of selection are given. If any unit is selected more than once it should be included in the sample.

|  |  |  |  |
| --- | --- | --- | --- |
| Random number | Sr. number of villages | Total population | Probability of Selection |
| 859677 | 483 | 7346 | .005946 |
| 74835 | 50 | 9231 | .006511 |
| 491741 | 275 | 3713 | .001335 |
| 285996 | 131 | 2310 | .001337 |
| 252541 | 108 | 7261 | .006127 |
| 287850 | 133 | 10425 | .00353 |
| 847258 | 478 | 6978 | .006409 |
| 410596 | 221 | 399 | .000316 |
| 674344 | 397 | 737 | .002414 |
| 727666 | 423 | 3203 | .001396 |
| 920794 | 508 | 4039 | .002813 |
| 291874 | 135 | 5439 | .000906 |
| 742201 | 434 | 1373 | .000885 |
| 37860 | 33 | 8074 | .006968 |
| 750855 | 437 | 3416 | .00166 |
| 91613 | 54 | 5841 | .003874 |
| 757074 | 441 | 1316 | .002297 |
| 213334 | 92 | 6475 | .004451 |
| 656265 | 385 | 1261 | .002064 |
| 843800 | 478 | 6975 | .006409 |
| 464793 | 258 | 2513 | .002781 |
| 598479 | 360 | 3039 | .001128 |
| 314161 | 153 | 322 | .000697 |
| 820668 | 472 | 13056 | .00613 |
| 18504 | 19 | 593 | .000998 |
| 32315 | 28 | 2515 | .001936 |

* 1. **EXPECTATION.**

If the ith unit is selected from a population of N units with probability , than an unbiased estimator,  of population total Y as suggested by Hansen and Hurwitz (1943) is

 **(7.3.1)**

where *HH* denotes the Hansen and Hurwitz, and *PPS* denotes probability proportional to size.

***THEOREM (7.1)***

If a sample of size n is drawn from a population of N units with probability proportional to size and with replacement then  is an unbiased estimator of population total, Y.

***PROOF.***

We know that

 **(7.3.1)**

Taking the expectation



**7.4. VARIANCE AND UNBIASED VARIANCE ESTIMATOR OF**

 **PROBABILITY PROPORTIONAL TO SIZE SAMPLING AND**

 **WITH REPLACEMENT**

***THEOREM (7.2)***

If a sample of size n is drawn from a population of N units with probability proportional to size and with replacement, the variance of  is

 **(7.4.1)**

***PROOF.***

We know that the variance of an estimator  is



Using (7.3.1), we have

 





Since the selection is independent; therefore Pij = PiPj,



on simplification we get



This expression may alternatively be written as

 **(7.4.2)**

 **(7.4.3)**

 **(7.4.4)**

***ALTERNATIVE PROOF (Using Indicator Variable)***

Let ai is defined as the number of times that the ith unit of the population is in the sample, then the joint distribution of ai is

 **(7.4.5)**

Then

 **(7.4.6)**

An unbiased estimator of population total is

 **(7.4.7)**

The unbiasedness can be proved easily as:-

Taking the expectation of (7.4.7) and putting E(ai) = nPi from (7.4.6) we get



The variance of  is

 **(7.4.8)**

Putting the values of Var(ai) and Cov (ai, aj) from (7.4.6) in (7.4.8) and on simplification the required variance formulas are obtained.

It follows that, if  the variance is zero. In practice, this ideal situation can of course not be realized as the probabilities cannot be chosen proportional to yi, which still has to be observed. But this situation can be approximated if it is possible to choose Pi proportional to some measures of size Zi which is known for all units in the population and which may be assumed approximately proportional to Yi. The Zi will then be called the size of the ith unit and least possible variance may be obtained by choosing the probabilities proportional to the sizes.

An analogous expression for the covariance of unbiased estimator  and  in the case of sampling with replacement and with probabilities proportional to size may be written in a straight far ward manner, i.e.

 **(7.4.9)**

* + 1. **Unbiased Variance Estimator**

***THEOREM (7.3)***

If a sample of size n is drawn from a population of N units with Probability Proportioned to size and with replacement then an unbiased variance estimator of (7.4.1) is

 **(7.4.10)**

***PROOF.***

Taking expectation of (7.4.10) 

By simple algebraic manipulation (7.4.10) becomes



Taking the expectation of the above equation



 

using (7.4.2) we get



using this result in (7.4.10), we get



(7.4.10) may be written as

 **(7.4.11)**

For calculation purpose alternative form of (7.4.10) is

 **(7.4.12)**

An unbiased covariance expression may be written analogies to (7.4.9) as

 **(7.4.13)**

Though this scheme is based on with replacement process but for the following reasons, it is preferred to be used in large scale sample surveys;

1. selection of the sample is simple,
2. can be used for any finite predetermined number of units in the sample,
3. an unbiased variance estimator is simple, and
4. it is also comparatively easy to obtain unbiased variance estimator of total in multistage designs.

This selection procedure may be more efficient than simple random sampling if the measure of size is approximately proportional to variable under study i.e. Yi and Zi are linearly related and regression line passing through the origin.

***Example (7.2.)***

Draw all possible samples of size 2 using Hansen and Hurwitz sampling procedure from the following data and show that = Y. Find the  and verify it by using the formula given (7.4.1).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Yi | 0.5 | 1.2 | 2.1 | 3.2 |
| Zi | 1 | 2 | 3 | 4 |

### SOLUTION

|  |  |  |
| --- | --- | --- |
| Yi | Zi |  |
| 0.5 | 1 | 0.1 |
| 1.2 | 2 | 0.2 |
| 2.1 | 3 | 0.3 |
| 3.2 | 4 | 0.4 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Possible Samples | yI | yj | pi | pj |  |  |  |
| 1 | 1 | 0.5 | 0.5 | .1 | .1 | 5.0 | 0.05 | 0.250 |
| 1 | 2 | 0.5 | 1.2 | .1 | .2 | 5.5 | 0.11 | 0.605 |
| 1 | 3 | 0.5 | 2.1 | .1 | .3 | 6.0 | 0.18 | 1.080 |
| 1 | 4 | 0.5 | 3.2 | .1 | .4 | 6.5 | 0.26 | 1.690 |
| 2 | 1 | 1.2 | 0.5 | .2 | .1 | 5.5 | 0.11 | 0.605 |
| 2 | 2 | 1.2 | 1.2 | .2 | .2 | 6.0 | 0.24 | 1.440 |
| 2 | 3 | 1.2 | 2.1 | .2 | .3 | 6.5 | 0.39 | 2.535 |
| 2 | 4 | 1.2 | 3.2 | .2 | .4 | 7.0 | 0.56 | 3.920 |
| 3 | 1 | 2.1 | 0.5 | .3 | .1 | 6.0 | 0.18 | 1.080 |
| 3 | 2 | 2.1 | 1.2 | .3 | .2 | 6.5 | 0.39 | 2.535 |
| 3 | 3 | 2.1 | 2.1 | .3 | .3 | 7.0 | 0.63 | 4.410 |
| 3 | 4 | 2.1 | 3.2 | .3 | .4 | 7.5 | 0.90 | 6.750 |
| 4 | 1 | 3.2 | 0.5 | .4 | .1 | 6.5 | 0.26 | 1.690 |
| 4 | 2 | 3.2 | 1.2 | .4 | .2 | 7.0 | 0.56 | 3.920 |
| 4 | 3 | 3.2 | 2.1 | .4 | .3 | 7.5 | 0.90 | 6.750 |
| 4 | 4 | 3.2 | 3.2 | .4 | .4 | 8.0 | 1.28 | 10.240 |
|  |  |  |  |  |  |  | 7.00 | 49.500 |



and

 = 49.50 – 49 = 0.50

Using formula

 

***Example (7.3)***

From Example (7.1) select a sample of 26 villages using probability proportional to size and with replacement selection procedure. Estimate the total number of trees in523 villages and compare this result with actual number of population given in 523 villages. Estimate and calculate standard error of this estimate.

***Solution:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No. of**Villages | **No. of Trees (yi)** | **Probability of Selection (pj)** |  |  |  |
| 84161110 | 3119491179924833044 | 0.0140.0360.2750.1210.212 | 22214.28626361.11142905.45520520.66114358.490 | 9349614.911186162.77310938735.2022575222.29119104700.50 | 6908642.925016694.4506241458.150952801.643707245.28 |
|  |  |  | **126360.003** | **463154435.5** | **632826842.1** |

(i) Estimated Total 



Actual Total 

(ii) 



(iii) 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No.** | **y** | **p** | **yi/pi** | **(yi/pi-y)^2** |
| 1 | 7346 | 0.005946 | 1235452.405 | 137886694865.41 |
| 2 | 9231 | 0.006511 | 1417754.569 | 35731892260.18 |
| 3 | 3713 | 0.001335 | 2781273.408 | 1379426820568.15 |
| 4 | 2310 | 0.001337 | 1727748.691 | 14632605885.20 |
| 5 | 7261 | 0.006127 | 1185082.422 | 177831700028.54 |
| 6 | 10425 | 0.00353 | 2953257.79 | 1812993331051.22 |
| 7 | 6978 | 0.006409 | 1088781.401 | 268326052675.04 |
| 8 | 399 | 0.000316 | 1262658.228 | 118422122062.40 |
| 9 | 737 | 0.002414 | 305302.4027 | 1693852740901.42 |
| 10 | 3203 | 0.001396 | 2294412.607 | 472833951008.72 |
| 11 | 4039 | 0.002813 | 1435833.63 | 29223818023.56 |
| 12 | 5439 | 0.000906 | 6003311.258 | 19329457362517.20 |
| 13 | 1373 | 0.000885 | 1551412.429 | 3065942449.29 |
| 14 | 8074 | 0.006968 | 1158725.603 | 200755773987.28 |
| 15 | 3416 | 0.00166 | 2057831.325 | 203444246707.55 |
| 16 | 5841 | 0.003874 | 1507743.934 | 9808812374.83 |
| 17 | 1316 | 0.002297 | 572921.2016 | 1068871009157.62 |
| 18 | 6475 | 0.004451 | 1454729.274 | 23120451813.51 |
| 19 | 1261 | 0.002064 | 610949.6124 | 991684897660.22 |
| 20 | 6975 | 0.006409 | 1088313.309 | 268811216688.66 |
| 21 | 2513 | 0.002781 | 903631.7871 | 494422166072.03 |
| 22 | 3039 | 0.001128 | 2694148.936 | 1182363847314.18 |
| 23 | 322 | 0.000697 | 461979.9139 | 1310574981674.20 |
| 24 | 13056 | 0.00613 | 2129853.181 | 273602014185.76 |
| 25 | 593 | 0.000998 | 594188.3768 | 1025348645657.47 |
| 26 | 2515 | 0.001936 | 1299070.248 | 94687373182.90 |
| **Total** | **117850** | **0.081318** | **41776367.9425** | **32621180470772.50** |
|    | **y'=** | **1606783.382** |    |
| **var(y')=** | **50186431493** |
| **S.E(y')=** | **224023.2834** |

|  |  |
| --- | --- |
| **yi^2/pi^2** | **yi^2/pi** |
| 1.52634E+12 | 9075633367 |
| 2.01003E+12 | 13087292428 |
| 7.73548E+12 | 10326868165 |
| 2.98512E+12 | 3991099476 |
| 1.40442E+12 | 8604883467 |
| 8.72173E+12 | 30787712465 |
| 1.18544E+12 | 7597516617 |
| 1.59431E+12 | 503800632.9 |
| 93209557065 | 225007870.8 |
| 5.26433E+12 | 7349003582 |
| 2.06162E+12 | 5799332030 |
| 3.60397E+13 | 32652009934 |
| 2.40688E+12 | 2130089266 |
| 1.34265E+12 | 9355550517 |
| 4.23467E+12 | 7029551807 |
| 2.27329E+12 | 8806732318 |
| 3.28239E+11 | 753964301.3 |
| 2.11624E+12 | 9419372051 |
| 3.73259E+11 | 770407461.2 |
| 1.18443E+12 | 7590985333 |
| 8.1655E+11 | 2270826681 |
| 7.25844E+12 | 8187518617 |
| 2.13425E+11 | 148757532.3 |
| 4.53627E+12 | 27807363132 |
| 3.5306E+11 | 352353707.4 |
| 1.68758E+12 | 3267161674 |
| **99746754265786.8** | **217890794431.5760** |

(i) Estimated Total 

****

(ii) 



(iii) 

This may be calculated as:

= 

 = 

 = 50186431493

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **y** | **pi** | **yi/pi** | **(yi/pi-y)^2** | **yi2/pi2** | **yi2/pi** |
| 7346 | 0.005946 | 1235452.405 | 137886694865.41 | 1526342644966.26 | 9075633366.97 |
| 9231 | 0.006511 | 1417754.569 | 35731892260.18 | 2010028018460.83 | 13087292428.20 |
| 3713 | 0.001335 | 2781273.408 | 1379426820568.15 | 7735481771381.28 | 10326868164.79 |
| 2310 | 0.001337 | 1727748.691 | 14632605885.20 | 2985115539595.95 | 3991099476.44 |
| 7261 | 0.006127 | 1185082.422 | 177831700028.54 | 1404420347090.44 | 8604883466.62 |
| 10425 | 0.00353 | 2953257.79 | 1812993331051.22 | 8721731576370.89 | 30787712464.59 |
| 6978 | 0.006409 | 1088781.401 | 268326052675.04 | 1185444939500.23 | 7597516617.26 |
| 399 | 0.000316 | 1262658.228 | 118422122062.40 | 1594305800352.51 | 503800632.91 |
| 737 | 0.002414 | 305302.4027 | 1693852740901.42 | 93209557064.60 | 225007870.75 |
| 3203 | 0.001396 | 2294412.607 | 472833951008.72 | 5264329213224.85 | 7349003581.66 |
| 4039 | 0.002813 | 1435833.63 | 29223818023.56 | 2061618211824.16 | 5799332029.86 |
| 5439 | 0.000906 | 6003311.258 | 19329457362517.20 | 36039746063769.10 | 32652009933.77 |
| 1373 | 0.000885 | 1551412.429 | 3065942449.29 | 2406880526030.20 | 2130089265.54 |
| 8074 | 0.006968 | 1158725.603 | 200755773987.28 | 1342645022480.99 | 9355550516.65 |
| 3416 | 0.00166 | 2057831.325 | 203444246707.55 | 4234669763390.91 | 7029551807.23 |
| 5841 | 0.003874 | 1507743.934 | 9808812374.83 | 2273291770267.82 | 8806732318.02 |
| 1316 | 0.002297 | 572921.2016 | 1068871009157.62 | 328238703205.28 | 753964301.26 |
| 6475 | 0.004451 | 1454729.274 | 23120451813.51 | 2116237261564.69 | 9419372051.22 |
| 1261 | 0.002064 | 610949.6124 | 991684897660.22 | 373259428895.50 | 770407461.24 |
| 6975 | 0.006409 | 1088313.309 | 268811216688.66 | 1184425859436.00 | 7590985333.13 |
| 2513 | 0.002781 | 903631.7871 | 494422166072.03 | 816550406706.21 | 2270826681.05 |
| 3039 | 0.001128 | 2694148.936 | 1182363847314.18 | 7258438490267.09 | 8187518617.02 |
| 322 | 0.000697 | 461979.9139 | 1310574981674.20 | 213425440862.56 | 148757532.28 |
| 13056 | 0.00613 | 2129853.181 | 273602014185.76 | 4536274572942.42 | 27807363132.14 |
| 593 | 0.000998 | 594188.3768 | 1025348645657.47 | 353059827068.97 | 352353707.41 |
| 2515 | 0.001936 | 1299070.248 | 94687373182.90 | 1687583509067.00 | 3267161673.55 |
|   | **y'=** | **1606783.382** | 32621180470772.50 | 99746754265786.80 | 217890794431.58 |
|   | **var(y')=** | **50186431493** |   |  |  |
|   | **S.E(y')=** | **224023.2834** |   |  |  |

**7.4.2. Comparison of Simple Random Sampling with Replacement**

 **and Probability Proportional to Size with Replacement**

 We know that

 **(7.4.1)**

If Pi =1/N then (7.4.1) becomes

 **(7.4.14)**

which is a variance expression for simple random sampling with replacement.

Putting Pi = Zi/Z in (7.4.1) and subtracting from (7.4.14), we obtain

 **(7.4.15)**

where .

Probability Proportional to size (PPS) sampling with replacement will be more efficient than simple random sampling provided.

  **(7.14.16)**

i.e. If Zi and  are positively correlated.

However, it was noted by Raj (1954) that estimator based on PPS sampling with replacement turns out to be inefficient compared to simple unbiased estimate based on simple random sampling with replacement if the regression line Yi on Zi is far from the origin.

* + 1. **Comparison of** **and** **Using a Linear Stochastic Model**

We have already shown that

  **(7.4.15)**

 

For the purpose of comparison, let us take the linear model as defined in the Chapter 6, i.e.

 Yi = α + βZi + εi, **(7.4.16)**

where E(εi) = 0 and Var(εi) = σ2 Zi2γ

E(eiεj) = 0 ½ ≤ γ ≤ 1

Substituting the value of Yi from the model in (7.4.15), we have



Using the condition of the model

 









We conclude that PPS sampling with replacement is more efficient as compared to simple random sampling, if



or



If Yi and Zi are perfectly linear i.e.



Then



which may not be satisfied all the times, which means that pps sampling is not necessary always more efficient. We, therefore, reach to a conclusion that the linearity of regression is not sufficient for pps sampling to be more efficient than simple random sampling. Further, if the line passes through the origin than

 Yi = βZi + εi

Under this situation



This satisfied only if γ ≥ ½, since σ2 ≥ 0 and 

***Example (7.4)***

Assuming that the finite population Y1, Y2, …., YN is a random sample from an infinite superpopulation in which 



Then show that

 

and



where



Hence prove that pps estimator is superior to equal probability (random sampling with replacement) if



# Solution

Given 

Summing over i we have



We know that



Putting the value of Yi and Y from the model, taking expectation and applying the conditions of model we have







Since 



Now





Putting the value Yi and Y from the model, taking expectation and applying the condition of model we have





Since  we have



Comparing 





We conclude that PPS estimator will be superior to equal probability if



or



* + 1. **Gain due to PPS sampling (with replacement) over simple random sampling**

We know that

 **(2.5.2)**

and

 **(7.4.12)**

We can prove that

(i)  **(7.4.16)**

and

(ii) 



 **(7.4.17)**

Using (7.4.16) and (7.4.17) in (2.4.7) we can have

 **(7.4.18)**

If f.p.c. is ignored we then have



 **(7.4.19)**

From (7.4.12) and (7.4.19), the gain due to pps sampling with replacement may be estimated as



On simplification we get

 **(7.4.20)**

An estimate of the percentage gain in efficiency due to pps sampling is

 **(7.4.21)**

***Example (7.5)***

From Example 7.3 estimate the gain of PPS sampling with simple random sampling.

***Solution***

We know from Example 7.3, know that  may be calculated using (7.4.19) which is

 **(7.4.19)**



= 383158187.5





* + 1. **Ratio Estimate for PPS Sampling with Replacement**

We know that



Therefore

 **(7.4.22)**

From Hansen, Hurwitz and Madow (1953), we have

 **(7.2.23)**

Using (7.4.22) and (7.4.9) and analogues expression

, **(7.4.24)**

in (7.2.23) and on simplification

.

. **(7.4.25)**

which is identical with (7.4.4).

This may be put easily as

. **(7.4.26)**

An approximate unbiased estimator of  may be written in a straight forward way [or may be derived from (6.2.19)]

, **(7.4.27)**

or

. **(7.4.28)**

* 1. **Lahiri’s Selection Procedure**

An other selection procedure has been suggested by Lahiri (1951) which does not require the cumulation of the measures of sizes. For this, suppose population of N units with measures of sizes Zi is given with Zmax be the maximum size among the N units. A pair of random numbers is chosen; one from 1 to N (say ith) and other 1 to Zmax (say R); if R exceeds the size of the ith unit; then that unit is rejected otherwise it is accepted. For selecting a sample of n units the process is repeated again and again till the required size is obtained. To prove that under this scheme the probability that ith unit is in the sample is proportional to Pi we proceed as follows:

The probability of selecting the ith unit at the first draw is  and the probability that the draw is not effective will be;



Hence the probability that the ith unit is selected at the second draw while the first is ineffective; 

If the process is continued in a similar way the probability of the ith unit in first effective is

P(ith) = PI (ith) + PII (ith) + ……..



. (7.8.1)

Since  is convergent. Hence

.(7.8.2)

Hence the probability that ith unit is in the sample is proportional to Pi.

**Example 7..4**

From the following data select a sample of size 2 using Lahiri’s procedure.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No. of****Village** | Area in Acres**Zi** | **Effective****Range** | **In-effective****Range** |
|  | 33 | 1 – 33 | 34 – 123 |
|  | 8 | 1 – 8 | 9 – 123 |
|  | 1 | 1 | 2 – 123 |
|  | 16 | 1 – 16 | 17 – 123 |
|  | 43 | 1 – 43 | 44 – 123 |
|  | 40 | 1 – 40 | 41 – 123 |
|  | 9 | 1 – 9 | 10 – 123 |
|  | 6 | 1 – 6 | 7 – 123 |
|  | 5 | 1 – 5 | 6 – 123 |
|  | 95 | 1 – 95 | 96 – 123 |
|  | 54 | 1 – 54 | 55 – 123 |
|  | 1 | 1 | 2 – 123 |
|  | 1 | 1 | 2– 123 |
|  | 2 | 1 – 2 | 3 – 123 |
|  | 1 | 1 | 2 – 123 |
|  | 123 | 1 – 123 | - |
|  | 1 | 1 – 1 | 2 – 123 |
|  | 3 | 1 – 3 | 4 – 123 |
|  | 4 | 1 – 4 | 5 – 123 |
|  | 2 | 1 – 2 | 3 – 123 |

**Solution**

Suppose a pair of random number selected (11, 68), the size of the 11th number is less than 68 so this selection is ineffective, we will proceed to select another pair of random number say (5, 38) the size of the 5th unit is 43 which is more than 38 so this draw is effective.

**Example (7.5)**

1. Select a sample size 5 villages using Lahiri’s method; Estimate the total number of trees and also estimate the standard error of the total.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No. of****Village** | **No. of Trees** **Yi** | Area in Acres**Zi** |  | **Effective****Range** | **In-effective****Range** |
|  | 2328 | 328 | 0.700 | 1 – 328 | 329 – 1234 |
|  | 754 | 80 | 0.020 | 1 – 80 | 81 – 1234 |
|  | 105 | 6 | 0.001 | 1 – 6 | 7 – 1234 |
|  | 949 | 156 | 0.030 | 1 – 156 | 157 – 1234 |
|  | 3091 | 428 | 0.100 | 1 – 428 | 429 – 1234 |
|  | 1736 | 401 | 0.090 | 1 – 401 | 402 – 1234 |
|  | 840 | 94 | 0.020 | 1 – 94 | 95 – 1234 |
|  | 311 | 63 | 0.010 | 1 – 63 | 64 – 1234 |
|  | 0 | 51 | 0.010 | 1 – 51 | 52 – 1234 |
|  | 3044 | 946 | 0.210 | 1 – 946 | 947 – 1234 |
|  | 2483 | 537 | 0.120 | 1 – 537 | 538 – 1234 |
|  | 128 | 7 | 0.002 | 1 – 7 | 8 – 1234 |
|  | 102 | 8 | 0.002 | 1 – 8 | 9 – 1234 |
|  | 60 | 22 | 0.005 | 1 – 22 | 23 – 1234 |
|  | 0 | 4 | 0.001 | 1 – 4 | 5 – 1234 |
|  | 11799 | 1234 | 0.280 |  1 – 1234 |  – |
|  | 26 | 3 | 0.001 | 1 – 3 | 4 – 1234 |
|  | 317 | 30 | 0.010 | 1 – 30 | 31 – 1234 |
|  | 190 | 40 | 0.010 | 1 – 40 | 41 – 1234 |
|  | 180 | 20 | 0.004 | 1 – 20 | 21 – 1234 |

# Solution

Selection of 5 villages with probability proportional to size using Lahiri’ Method with replacement is as:-

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pair of Random Numbers** | **No. of****Village** | **Yi** | **Pi** |  |  |
| (16,37) | 16 | 11799 | 0.28 | 42139.286 | 1.7757194 x 109 |
| (5,38) | 5 | 3091 | 0.10 | 30910.00 | 9.554281 x 109 |
| (16,23) | 16 | 11799 | 0.28 | 42139.286 | 1.7757194 x 109 |
| (10,54) | 10 | 3044 | 0.21 | 14495.238 | 2.1011192 x 108 |
| (10,54) | 10 | 3044 | 0.21 | 14495.238 | 2.1011192 x 108 |
|  |  |  |  | 144179.05 | 4.9270908 x 109 |

An unbiased estimate of population total



Therefore the total number of trees is 28836.

 

= 38478561



**Example 7.5**

Following is an other example to explain how a sample of size two is selected using Lahiri’s selection procedure.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Hours | **1** | **2** | **3** | **4** | **5** | **6** |
| Number of Boats Zi | 42 | 52 | 19 | 6 | 23 | 56 |
| Catch of Fish in Kg | 568 | 887 | 223 | 88 | 352 | 1295 |
| Pi | 0.128 | 0.158 | 0.058 | 0.018 | 0.070 | 0.170 |
|  |  |  |  |  |  |  |
| **Hours** | **7** | **8** | **9** | **10** | **11** | **12** |  |
| Number of Boats Zi | 36 | 59 | 14 | 14 | 2 | 6 | 329 |
| Catch of Fish in Kg | 934 | 1265 | 486 | 443 | 98 | 0 |  |
| Pi | 0.170 | 0.179 | 0.043 | 0.043 | 0.006 | 0.018 |  |

**Solution**

Using Lahiri’s selection procedure a pair of random number is selected between [1 – 12 and 1 – 59] since the maximum size is 59. The pair of random number is [2, 38] since 38 is less than the size of the 2nd unit, so 2nd unit is in the sample with probability 52/329 = 0.158 with Yi = 0.158.

Since the selection is without replacement, therefore, an other unit will be selected from the remaining 11 units. They are put in the following Table:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hours** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
| No. of Boats Zi | 42 | 19 | 6 | 23 | 56 | 36 | 59 | 14 | 14 | 2 | 6 |
| Yi | 568 | 223 | 88 | 352 | 1295 | 934 | 1265 | 486 | 443 | 98 | 0 |

Again a pair random number is selected on this time between [1 – 11 and 1 – 59], the selected random number pair is [6, 32]. Since the size of the 6th unit in the above table is greater than 32 so the 6th unit of the second table [which is infact the 7th unit in the original table is in the sample] with probability 0.170 and Yi = 934.

**7.6 SAMPLING WITH UNEQUAL PROBABILITIES**

**WITHOUT REPLACEMENT (πPS)**

Sampling with unequal probabilities without replacement is not so simple as compared to with replacement procedure. For that let a sample of two units is selected from a population of N units. Let the probability of the selection of the ith unit is Pi = Zi/Z. Suppose the ith unit is not selected at the first draw but the jth unit is selected (j ≠ i) then the probability of selecting the jth unit at the first draw is Pj = Zj/Z; and the conditional probability of selecting the ith unit at the second draw is Zi/(Z - Zi) = Pi/(1 - Pj). The probability of inclusion of ith unit at the second draw to be included in the sample is the sum of the product that the jth unit is selected at the first draw and the ith unit is selected at the second draw given the jth unit is selected at the first draw i.e.

 **(7.6.1)**

the total probability πi, the probability of inclusion of the ith population unit to be in the sample is

  **(7.6.2)**

The probability that both ith and jth units are in the sample is denoted by πij and is defined as

|  |  |  |
| --- | --- | --- |
| πij | = | Probability of the selection of the ith unit × probability of selection of jth unit given the ith unit is selected at the first draw + probability of the selection of jth unit × probability of the ith unit given jth unit is selected at the first draw |



 **(7.6.3)**

For any sample size similar expressions may be obtained but as the sample size increase the expressions are becoming more complicated.

If a population of 4 units is given with the probabilities 0.1, 0.2, 0.3, and 0.4; for n = 2; π1 = 0.2345; π2 = 0.4413; π3 = 0.6083 and π4 == 0.7159 and π12 = 0.0472; π13 = 0.0762; π14 = 0.1111; π23 = 0.1607; π24 = 0.2333; π34 = 0.3714.

**7.6 HORVITZ AND THOMPSON ESTIMATOR**

The general theory of unequal probabilities sampling without replacement was presented firstly by Horvitz and Thompson (1952). An unbiased estimator suggested by them for population total Y is

 **(7.7.1)**

where ai = 1 if the ith unit is in the sample 0 otherwise

then ai follows the binomial distribution with sample size 1 with probability πi.

 **(7.7.2)**

(There are n values of ai and N-n values are zero)

E(a­i) = πi; E(ai,aj) = πij **(7.7.3)**

Var(ai) = πi (1-πi) **(7.7.4)**

Cov (ai,aj) = πij - πiπj **(7.7.5)**

 **(7.7.6)**

 **(7.7.7)**

 **(7.7.8)**

The proof of the three relations is very simple

(i) ****

Taking the expectation of (7.7.2) and using (7.7.3) we get the relation. Alternatively; Σπi is the sum of the probabilities of all samples containing ith unit, of all samples containing the jth unit and so on and for the totality of sample s of size n, ΣP(s) = 1. Thus every P(s) occurs n times in this sum, once as a sample containing the first unit in it, then as a sample containing the second unit in it and so on. Hence Σπi = n.

(ii) 

 is sum of all the probabilities of the samples containing ith and jth units and  is the sum of the probabilities containing first and second units; first and third units; and so on. Thus every P(s) containing the first unit occurs (n-1) times in this sum as the sample has (n-1) other members in it and it occurs once for each of these members. Hence  and so on. In general .

 (iii) 

Taking the left hand side

 



Substituting (7.7.7) in the above expression

= (n – 1) πi - πi (n - πi) = -πi(1-πi)

The following useful result is also presented

(iv) 

 **(7.7.9)**

**7.8 EXPECTATION, VARIANCE AND UNBIASED VARIANCE**

ESTIMATOR OF HORVITZ AND THOMPSON ESTIMATOR

***THEOREM (7.4)***

If a sample of n units is drawn without replacement and with probability proportional to size, an unbiased estimator of population total Y is

 **(7.7.1)**

###### PROOF

Taking the expectation of (7.7.1)



Using (7.7.3) i.e. E(ai) = πi



 **THEOREM (7.5)**

The variance of  is

 **(7.8.1)**

# PROOF







 **(7.8.2)**

Substituting the values of Var(ai) and Cov(ai,aj) from (7.7.4) and (7.7.5) in (7.8.2)we get

 **(7.8.1)**

Since

 **(7.7.7)**

Substituting 1 - πi from (7.7.7) in the leading term of the right hand side of expression (7.8.1)







 **(7.8.3)**

This expression was put forward by Yates and Grundy (1953) and Sen (1953) independently which is valid if the sample size is fixed. This is known as Sen-Yates-Grundy variance expression.

The following are unbiased variance estimators of (7.8.1) suggested by Horvitz and Thompson (1952)

 **(7.8.4)**

This estimator suffers from the disadvantage that it is not always zero when the variance is zero. The following alternative unbiased variance estimator was derived by Sen (1953) and by Yates and Grundy (1953) for use when the number of sample units are fixed

 **(7.8.5)**

provided πij is not equal to zero for i ≠ j i.e. if all the possible pairs of distinct population units have non zero probability of inclusion in the sample. This is known as Sen-Yates-Grundy variance estimator. Both these estimators can assume negative values but (7.8.5) rarely seems to do so in practice. (7.8.5) is always non negative [Sen (1953), Raj (1956), Brewer (1963)]. Rao (1961, 1963) obtained the same result under two well-known selection procedures for unequal probabilities without replacement. Rao and Sing (1973) compared (7.8.4) and (7.8.5) for 34 populations using Brewer’s selection procedure for n = 2 and came to the conclusion that (7.8.5) is always non-negative whereas (7.8.4) is negative for some of the pairs. Both these estimators can assume negative values but (7.8.5) rarely seems to do so in practice also for most of the well-known selection procedures for unequal probabilities without replacement (7.8.5) is always non negative moreover (7.8.5) has performed much better than (7.8.4) in worked examples as it leads to zero variance when Yi is proportional to πi for all i [Brewer and Hanif, 1969a]. For n = 2 it is the only possible non-negative variance estimator [Vijayan, 1975]. Similar result was obtained by Lanke (1974). Rao (1979) has shown that for n = 2 (7.8.5) is the unique unbiased variance estimator. The important properties of Horvitz and Thompson estimator are as;

1. It is the only unbiased estimator of the class in which same weight is attached to a particular population unit whenever it is selected (Horvitz and Thompson, 1952).
2. It is admissible in the class of all homogeneous linear unbiased estimators of population total Y, that is, there does not exist any member of that class which has smaller variance than  (Roy and Chakravarti, 1960, Godambe, 1960).
3. If Yi are exactly proportional to πi and the number of units in the sample is fixed the variance of  is zero, this is a property usually associated with ratio estimator and will be referred to as the ratio estimator property.
4. Under the model the expected variance of the Horvitz-Thompson estimator achieves the lower bound of the expected variance for any design-unbiased estimator (Godambe-Joshi, 1965).

If follows from ratio estimator property if the values of the measure of size are known for all units in the population and Yi are approximately proportional to the measure of size the variance of  can be made small by setting the proportional to πi. This is a principal reason why selection with probability proportional to size has assumed importance in unequal probability.

### Example 7.7

From the population given in Example 7.2, draw all possible samples of size 2, using the following selection procedure, i.e.

1. The first unit is selected with probability proportional to size Zi.
2. The second unit is selected with probability proportional to size from the remaining unit.

Show that the Horvitz-Thompson estimator is unbiased estimator of population total. Calculate the variance of this estimator.

# SOLUTION

Since sampling is without replacement the total probability of the inclusion of the ith unit at the first or at the second draw will be

 **(7.6.2)**

and the probability that the ith and jth units both included in the sample will be

 **(7.6.3)**

πi and πij have been calculated before. We will use these values and prove that Horvitz-Thompson estimator is unbiased estimator of population total.

The calculation, are given as

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **All Possible Samples** | **yi** | **yj** | **pj** | **pj** | **πij** |  |  |  |
| 1 | 2 | 0.5 | 1.2 | .1 | .2 | .0472 | 4.8514 | 0.2290 | 1.1109 |
| 1 | 3 | 0.5 | 2.1 | .1 | .3 | .0762 | 5.5844 | 0.4255 | 2.3763 |
| 1 | 4 | 0.5 | 3.2 | .1 | .4 | .1111 | 6.6021 | 0.7335 | 4.8426 |
| 2 | 3 | 1.2 | 2.1 | .2 | .3 | .1607 | 6.1715 | 0.9918 | 6.1206 |
| 2 | 4 | 1.2 | 3.2 | .2 | .4 | .2333 | 7.1891 | 1.6772 | 12.0577 |
| 3 | 4 | 2.1 | 3.2 | .3 | .4 | .3714 | 7.9221 | 2.9423 | 23.3080 |
|  |  |  |  |  |  |  |  | 6.9993 | 49.817 |

For n = 2 

(i)  = 6.9993, where

 is for all possible samples

(If the calculation are made more decimal places the result will be exactly 7)

(ii)  = 49.817 – 49 = 0.817

This may also be calculated directly using (7.8.7).

**7.8.1 Comparison of with and without Replacement**

**for unequal Probability Sampling**

We know that

 **(7.4.3)**

Substituting πi = nPi in (7.4.3)

 **(7.8.6)**

Comparing (7.8.3) and (7.8.6)  if

 **(7.8.7)**

Further comparison can also be made as:

 **(7.4.1)**

 **(7.4.4)**

and

 **(7.8.1)**

 **(7.8.9)**

If πi = n Pi  we have

 **(7.8.10)**

 **(7.8.11)**

From (7.4.4) and (7.8.11) we will have

 **(7.8.12)**

This shows that 

 **(7.8.13)**

**7.8.2. Ratio Estimator for unequal Probability and Without Replacement**

We know that



Then ratio estimator is defined as

 **(7.8.14)**

with

 **(7.8.15)**

where 

**7.9 PROPERTIES OF GOOD SELECTION PROCEDURE**

The desirable characteristics of selection procedure for sampling without replacement with probability proportional to size are as follows:

1. The number of units in the sample should be fixed.
2. The Pi should be precisely proportional to the measure of size.
3. Selection should be strictly without replacement.
4. There should be no difficulty in selecting a sample more than two units i.e. should be applicable for any n.
5. The joint probabilities of selection should be simple to calculate as they are required for estimation of variance.
6. Each pair of distinct population units should have a non-zero probability of selection i.e. Pij ≠ 0 for all j ≠ i.
7. The value of the Pij should be such as to minimize the variance of the variance estimator.
8. The selection procedure should be simple.
9. For any value of n, the Pi should take any value upto the theoretical limit n-1.
10. It should be simple to rotate. (Rotation means to drop some of the selected units and add new units in place of those).

**7.10 CLASSIFICATION BY MANNER OF SELECTION**

Since, 1949, it is not surprising that more than 130 selection procedures for sampling without replacement with unequal probabilities have been appeared in literature. All these selection procedures can be broadly put forward in four main classification by manner of selection. These classifications were set out by Carrol and Hartley (1964).

**(a) Draw by Draw Procedure**

At each draw one unit is selected from among those units not already selected. Probabilities of selection are defined from each draw (since the selection is without replacement) almost always depend on the units already selected. If the probability of selection at a given draw are (apart from a normalizing factor) independent of which units are selected at previous draw, these probabilities are sometimes called as ***WORKING PROBABILITIES.***

**(b) Systematic Procedure**

Systematic procedure involves an ordering of population units or cumulating of size measures. The order of units may or may not be random. A random real number r (0 < r < 1) is chosen and n units selected are those whose cumulated values of πi (the desired probabilities of inclusion) are equal or to next greater than r, r + 1, r + 2, …… + r + n – 1.

**(c) Rejective Procedure**

The term rejective has been employed by Hajek (1964) and is somewhat wider in its connotation than the term mass draw used by Carrol and Hartley (1964). Rejective procedure resembles draw by draw procedures in that only a single unit is selected at each draw of n successive draws. They differ from ordinary draw by procedures in that the selection at a given draw may give rise to the selection of an already selected unit, in such case the selected sample is abandoned and the selection recommenced.

**(d) Whole Sample Procedure**

In these selection procedures the units are not individually drawn: a probability is specified for each possible sample of n distinct units and one selection using these probabilities selects the whole sample.

**7.11 SELECTION PROCEDURES FOR SAMPLING WITH UNEQUAL**

 **PROBABILITIES WITHOUT REPLACEMENT**

 **[Methods Using the Horvitz-Thompson Estimator]**

In this Section, those selection procedure will be considered in which Horvitz and Thompson estimator is involved i.e.

1. Midzuno’s Selection Procedure.
2. Yates and Grundy Draw by Draw Procedure.
3. Brewer’s Procedure.
4. Durbin’s Procedure.
5. Rao-Sampford Procedure.
6. Random Systematic Procedure.
7. Samiuddin-Asad Procedure.

**(a) Midzuno’s Procedure (1950)**

In this selection procedure the first unit is selected with probability proportional to size, Zi and the remaining units are selected with equal probability without replacement. If from a population of N units a sample of 2 units is selected, the probabilities that the ith unit is included in the sample is

 **(7.11.1)**

and the probability that both ith and jth units are in the sample is

|  |  |  |
| --- | --- | --- |
| πij | = | ith unit is selected at the first draw and jth unit at any of the subsequent draw + jth unit is selected at the first draw and ith unit at any of the subsequent draw. Neither ith and nor jth units are selected at the first draw but both of them are selected from (n – 1) draw |

 **(7.11.2)**



Similarly πijk : probability of inclusion of ith, jth and kth units are



 **(7.11.3)**

Under this scheme of sampling Yates and Grundy variance estimator is always non-negative. The main advantage of this selection procedure is that it is simple to compute.

It is also interesting to note that the probability of getting a particular unordered sample is the sum of the probabilities of the units in the sample i.e.

 **(7.11.4)**

and with this property the classical ratio estimated as defined in Chapter 6 becomes unbiased.

This procedure break down unless  which is a very stringent requirement, consequently this procedure is not frequently applicable. Rao (1963) has shown that for n = 2, the variance of  with this procedure is always smaller than the variance of  provided  which is also condition for non-negativity of the working probabilities.

Example 7.8

From the following data, draw all possible samples of size 2 under Midzuno’s selection procedure show that under this selection procedure = Y. Also calculate the 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr.No. | 1 | 2 | 3 | 4 |
| Zi | 1 | 2 | 3 | 4 |
| Yi | 0.5 | 1.2 | 2.1 | 3.2 |

SOLUTION

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **All Possible Samples** | **yi** | **yj** | **pj** | **pj** | **πij** |  |  |  |
| 1 | 2 | 0.5 | 1.2 | 0.1 | 0.2 | 0.100 | 3.821 | 0.3821 | 1.460 |
| 1 | 3 | 0.5 | 2.1 | 0.1 | 0.3 | 0.133 | 5.188 | 0.692 | 3.590 |
| 1 | 4 | 0.5 | 3.2 | 0.1 | 0.4 | 0.167 | 6.583 | 1.099 | 7.235 |
| 2 | 3 | 1.2 | 2.1 | 0.2 | 0.3 | 0.167 | 6.509 | 1.087 | 7.155 |
| 2 | 4 | 1.2 | 3.2 | 0.2 | 0.4 | 0.200 | 7.905 | 1.581 | 12.498 |
| 3 | 4 | 2.1 | 3.2 | 0.3 | 0.4 | 0.233 | 9.271 | 2.160 | 20.025 |

π1 = 0.4000

π2 = 0.4667

π3 = 0.5333

π4 = 0.6000



= 51.883 – 49 = 2.883.

**(b) Yates and Grundy Draw Procedure (1953)**

In this selection procedure the first unit is selected with probability proportional to size, Zi and the second unit is selected from the remaining units with probability proportional to size. If from a population of N units a sample of 2 units are selected, the total probability that the ith unit is included in the sample is (explained before)

 **(7.6.2)**

It can be easily seen that πi ≠ nPi

The probability that both ith and jth units are in the sample is

 **(7.6.3)**

The selection may proceed to n = 3 or more but formulas for πi, πij and so on, become rapidly complicated.

**(c) Brewer’s Procedure (1963)**

In this selection procedure the first unit is selected with probability proportional to Pi(1 – Pi)/(1 – 2Pi) and the second unit is selected with probability proportional to Pj(j ≠ i) from the remaining units. For n = 2 it can be shown that for this selection procedure that πi = 2Pi as







 **(7.11.7)**

Now the denominator may be simplified as

 **(7.11.8)**

From (7.11.7) and (7.11.8), we get πi = 2Pi

The joint probabilities of inclusion of two units πij, may be obtained as





Using (7.11.8) in the denominator we get

 **(7.11.9)**

It can be easily seen that (πiπj - πij) > 0 for all j ≠ i so that Yates and Grundy variance estimator is always positive. The author himself in 1975 generalized this method for any sample size n. By this method, variance of  is always less than  [Brewer, 1963]. Moreover, (πiπj - πij) is always positive [Rao, 1965], hence Yates-Grundy variance estimators is always positive.

***EXAMPLE (7.9)***

For Brewer’s method with n = 2, we have



1. Show that if every pi < , 0 < πij < 4pipj for j ≠ i.
2. Show that this makes the Sen – Yates – Grundy variance estimator always positive for this selection procedure.

# SOLUTION

(a) (1 – pi – pj) = 1 – pi – p­j + 2pipj  - 2pipj.

 = 1 – 2pi – p­j + 2pipj + pi - 2pipj.

 = (1 – 2pi) – p­j (1 – 2pi) + pi  (1 – 2pj).

 = (1 – 2pi) (1 – 2j) + pi (1 – 2pj).

 = (1 – 2pi) (1 – 2pj) 

 = (1 – 2pi) (1 – 2pj) 

 = (1 – 2pi) (1 – 2pj) 

Now



Substituting the value of (1 – pi – pj) from the above relation we have







since



so



 ⇒ 

 ⇒ 

(b)  **(7.8.5)**



since 

therefore



 ⇒ 

This shows that Yen-Yates-Grundy variance estimator is always positive if for all values 

Hence .

**(d) Durbin’s Procedure (1967)**

In this selection procedure the first unit is selected with probability proportional to Pi and second unit is selected with probability proportional to  from the remaining units.

We can show that for this selection procedure with n = 2, πi = 2pi

In this selection procedure as we know that the first unit is selected with probability proportional to pi and the second unit with probability proportional to  from the remaining units: i.e.

|  |  |  |
| --- | --- | --- |
| πi | = | **P** [ith unit is selected at the first draw] + P [ith unit is selected at the second draw]. |
|  | = | **P** [ith unit is selected at the first draw] (that the ith unit is selected at the second draw) P(that the ith unit is selected at the 2nd draw given the jth unit is selected at the first draw)] |





since 

 

also



Then

 = 2pi

and  comes out to be

 **(7.11.10)**

Since the probability in i and j is symmetric hence equals the probability of drawing these units in the reverse order

 **(7.11.11)**

Since



 **(7.11.12)**



Using (7.11.11) and (7.11.12) we get

 **(7.11.9)**

which is the same as (7.11.9).

Since for n = 2, Durbin’s Selection Procedure and Brewer’s Selection Procedure are identical, here variance under this method is also less than .

It can be easily shown that (πiπj - πij) > 0 hence Yates Grundy variance estimator is always positive. This procedure can be extended for n > 2 in principle.

It follows that for n = 2

 **(7.11.11)**

which is less than 

We can Show that for Durbin’s selection procedure with n = 2, πi = 2pi

**(e) Rao-Sampford Procedure (Rao 1965, Sampford 1967)**

In this selection procedure the first unit is selected with probability proportional to size Pi and the second unit is selected with probability proportional to , selection is with replacement. If any unit is selected twice, the sample (as selected upon the point) is abandoned and selection process commenced afresh. The πij for n = 2 can be obtained as

The probability of the cases where i and j are equal



The probability of the case where i and j are not equal



where



The probability that the ith unit is selected first jth second is



The probability that the jth unit is selected first and ith second



Hence the joint probability that both ith and jth units are in the sample is





Putting the value of K we obtain



or



or



on simplification we get



which is the same as (7.11.9).

This procedure is applicable for any n. For n > 2, πij are unduly complex and cannot be handled easily.

(πiπj - πij) > 0 so Yates and Grundy variance estimators is always positive. Under this selection procedure the variance of Horvitz and Thompson estimator is always less than variance . This procedure was first suggested for n = 2 by Rao (1965) but Sampford (1967) generalized for any n. It can also be proved that πi = nPi provided nPi < 1 for all units in the population.

Since πij are same for Brewer’s, Durbin’s and Sampford’s procedures for n = 2 hence for calculation purpose they will be referred as BDS selection procedure.

Example (7.11)

From the data given in Example (7.2) draw all possible samples of size 2 using Brewer’s Selection Procedure i.e.

(i) The first unit is selected with probability proportional to Pi(1–Pi)/1-2 Pi).

(ii) Second unit is selected with probability proportional to Pj(j ≠ i) from the remaining units.

Prove that the Horvitz-Thompson estimator is an unbiased estimator of population total Y and calculate the variance of .

# SOLUTION

Since in this selection procedure πi = nPi, there is no need to calculate πi. The joint probability of the ith and jth unit both in the sample will be

 **(7.11.9)**

The calculation are as:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **All Possible Samples** | **yi** | **yj** | **pj** | **pj** | **πij** |  |  |  |
| 1 | 2 | 0.5 | 1.2 | .1 | .2 | .0277 | 5.5 | 0.1524 | .8382 |
| 1 | 3 | 0.5 | 2.1 | .1 | .3 | .0535 | 6.0 | 0.3210 | 1.9260 |
| 1 | 4 | 0.5 | 3.2 | .1 | .4 | .1188 | 6.5 | 0.7722 | 5.0193 |
| 2 | 3 | 1.2 | 2.1 | .2 | .3 | .1188 | 6.5 | 0.7722 | 5.0193 |
| 2 | 4 | 1.2 | 3.2 | .2 | .4 | .2535 | 7.0 | 1.7745 | 12.4215 |
| 3 | 4 | 2.1 | 3.2 | .3 | .4 | .4277 | 7.5 | 3.2077 | 24.0577 |
|  |  |  |  |  |  |  |  | 7.0000 | 49.2820 |

**(f) Random Systematic Procedure (good man and Kish 1949).**

Arrange the population units in random order. Cumulate the measure of size, divide the total measure of size Z by the required number of units in sample, n, to obtain the skip interval Z/n. Choose a random start that is a random number greater than or equal to zero and less than the skip interval. The first unit selected is that for which the cumulate size measure is the smallest greater than or equal to the random start the second unit is that for which the cumulated size measure is the smallest greater than or equal to the random start plus the skip interval.

For this type of selection procedure Hartley and Rao (1962) have given a formula for π*ii*, which is asymptotically correct as *N →*  ∞ under certain conditions.

The main drawbacks of the systematic procedures are the difficulty of calculating the joint probabilities of inclusion for the purpose of estimating the variance, and the fact that one or more of these joint probabilities is sometimes zero. A simple example of a situation in which one of the π*IJ* is zero is given by *n* = 2 ; *N* = 5 ; *Zij* = 1, 2, 4, 5, 6.

Let us have a population of 8 units arranged in random order with the sizes. Let a sample of 3 units is to be selected

|  |  |  |
| --- | --- | --- |
| Unit | Size | Cumulative Total |
|  | 15 | 15 |
|  | 81 | 96 |
|  | 26 | 112 |
|  | 42 | 164 |
|  | 20 | 184 |
|  | 16 | 200 |
|  | 45 | 245 |
|  | 55 | 300 |

In this example Z/n = 300/3 = 100. Let the random start is 36, so the second unit is in the sample. For the selection of second unit Z/n = 100 will be added in 36 and comes out to be 36 + 100 = 135; this falls against 164 so ‘th unit is selected. Similarly for the selection of third unit 2Z/n will be added in 36 and so on till the required sample is obtained. The selection procedure is simple. The only disadvantage with this is, that no exact formula for variance and variance estimator is available. Hartley and Rao (1962) have derived an expression for the joint probabilities πij which is asymptotically correct as N→ ∞ under certain conditions. The values of πij are substituted in (7.8.1) and following asymptote formula for variance of  was obtained

 **(7.11.14)**

 **(7.11.15)**

It was show that for n = 2

 **(7.11.16)**

and the variance estimator of the same order is0

 1 **(7.11.17)**

Note that npi < 1.

Both these expression in case of equal probability sampling, πi = n/N, reduced to the variance expression of simple random sampling without replacement.

If in (7.11.14) the term  is deleted it comes out (7.4.2) which is an variance estimator. The factor mentioned here can be treated as correction factors. Hence it is obvious that  for Hartley-Rao scheme is less than . Connor (1966) has derived an exact formula for the πij for any n and J.N.K. Rao (1965) derived asymptotic expression for πij for various selection procedures. Expression (7.11.14) is valid for all the selection procedures using Horvitz and Thompson estimator provided πi = nPi (Hanif, 1974).

# Example 12

From the population given in Example 3, n = 2, find the variance of  under random systematic selection procedure.

# SOLUTION

Since πi = nPi

 **(7.11.15)**

Substituting the value of Pi and Yi we get that



**7.12 SELECTION PROCEDURES SAMPLING WITH UNEQUAL**

 **PROBABILITIES WITHOUT REPLACEMENT**

 **[Methods Using Special Estimators]**

In this Section these selection procedures will be considered which do not involve the Horvitz-Thompson estimator i.e.

1. Raj’s estimator with Yates and Grundy raw by Draw Procedure.
2. Murthy’s Estimator with Yates and Grundy Draw by Draw Procedure.
3. Rao, Hartley and Cochran (RHC) Estimator with RHC Selection Procedure.
4. Lahiri’s Estimator with Lahiri’s Selection Procedure.
5. Poisson Sampling

**(a) Raj’s Estimator**

 **[Raj, (1956)]**

In this selection procedure the first unit ith (say) is selected with probability proportional to size p­i and the second unit jth (say) is selected from amongst the remaining units with probability proportional to pj(j ≠ i) and so on (Yates and Grundy draw by draw procedure). The probabilities of the successive units being drawn are pi, pj/(1 – pi), pk/(1 – pi – pj) and so on. If y1, y2, …., yn are the units selected in the sample of size n in the same order then unbiased estimators of population total Y are

 **(7.12.1)**

Since order of selection is taken into account while selection, hence such estimators are called **ORDERED** estimators. All these estimators are unbiased of population total Y and are uncorrelated. Any linear combination ΣCiti where ΣCi = 1 is also an unbiased estimator of Y. These ordered estimates were proposed first time by Das (1951).

**THEOREM (7.6)**

E(t1) = E(t2) = … = E(tn) = Y and Var(t2) < Var(t1)

PROOF















Similarly it can be proved that E(tn) = Y

We can easily prove that e(t1t2) = Y2 as:

E(t1t2) = E1[t1E2­ E(t2|t1) = Y2

 = E1[t1 Y] = E­1(t1) y = y 2



 **(7.12.2)**

and





Now, to obtain variance of this estimator we proceed as under:

 = 

 = 

 = 

 =

 = 

 = 

 = 

 = 

 = 

 = 

 

 = 

 

 = 

 

 = 

 

 = 

 

 = 

 

 = 

 

 = 

 

 = 

 

 = 

 = 

 =  **(7.12.3)**

It is obvious from (7.12.2) and (7.12.3) that Var(t2) < Var(t1). Similarly it can be proved that Var(tk) < Var(tk-1).

**THEOREM (7.7)**

An unbiased estimator tmean of population total Y suggested by Raj (1956) is for n = 2 is tmean = (t1 + t2)/2 which the mean of these t1 and t­2 is

tmean 

 **(7.11.4)**

with variance

 (**7.11.5)**

PROOF

Since

E(t1) = E(t2) = Y Therefore E(tmean) = Y

 We have

 

 

 Now

 

 

  as t1 and t2 are independent.

 

 

 

 

 

After some algebric manipulation and on simplification we get



This may be put as





which is obviously less than Var(t1) and Var (t2).

An unbiased estimator for n = 2 is

 **(7.12.6)**

The variance for the general case is complicated and was derived by Pathak (1967) for any n and is given as

 **(7.12.7)**

where Qij(r – 1) denotes the probability of non-inclusion of one or both of the units i and j in the first (r – 1) sample.

An unbiased variance estimator suggested by Raj (1956) for any n is

 **(7.12.8)**

which is non negative for 

# Example 7.13

From the population given in Example 7.3 draw all possible ordered sample of size 2 (Raj’s scheme) and show that E(tmean) = Y and calculate the variance of tmean.

# SOLUTION

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **All Possible Samples** | **yi** | **yj** | **pj** | **pj** |  |  |  |  |  |
| 1 | 2 | 0.5 | 1.2 | 0.1 | 0.2 | 5 | 5.9 | 4.45 | 0.1211 | 0.6600 |
| 1 | 3 | 0.5 | 2.1 | 0.1 | 0.3 | 5 | 5.6 | 5.90 | 0.1967 | 1.1605 |
| 1 | 4 | 0.5 | 3.2 | 0.1 | 0.4 | 5 | 7.7 | 6.35 | 0.2822 | 1.7920 |
| 2 | 1 | 1.2 | 0.5 | 0.2 | 0.1 | 6 | 5.2 | 5.60 | 0.1400 | 0.7840 |
| 2 | 3 | 1.2 | 2.1 | 0.2 | 0.3 | 6 | 6.8 | 6.40 | 0.4800 | 3.0720 |
| 2 | 4 | 1.2 | 3.2 | 0.2 | 0.4 | 6 | 7.6 | 6.80 | 0.6800 | 4.6240 |
| 3 | 1 | 2.1 | 0.5 | 0.3 | 0.1 | 7 | 5.6 | 6.30 | 0.2700 | 1.7010 |
| 3 | 2 | 2.1 | 1.2 | 0.3 | 0.2 | 7 | 6.3 | 6.65 | 0.5700 | 3.7905 |
| 3 | 4 | 2.1 | 3.2 | 0.3 | 0.4 | 7 | 7.7 | 7.35 | 1.2600 | 9.2610 |
| 4 | 1 | 3.2 | 0.5 | 0.4 | 0.1 | 8 | 6.2 | 7.10 | 0.4733 | 3.36.4 |
| 4 | 2 | 3.2 | 1.2 | 0.4 | 0.2 | 8 | 6.8 | 7.40 | 0.9867 | 7.3016 |
| 4 | 3 | 3.2 | 2.1 | 0.4 | 0.3 | 8 | 7.4 | 7.70 | 1.5400 | 11.8580 |
|  |  |  |  |  |  |  |  |  | 7.0000 | 49.365 |





 = 49.365 – 49 = 0.3365

**(b) Murthy’s Estimator**

 **[Murthy (1957)]**

Murthy (1957) used the same selection procedure as used by Raj (1956). The estimator given by Raj [discussed above] depend on the order in which the units are drawn from the population. Murthy has shown that the estimator tmean could be improved by the process of unordering. **Unordered estimator is that which does not depend on the order in which the units are drawn.** This improvement is in the sense that the variance of unordered estimator is less than variance of ordered estimator. Murthy himself proposed this estimator.

 **(7.12.9)**

where P(s) is the probability of getting sth unordered sample and  is the conditional probability that the sth sample was selected given that the ith unit has been selected at the first drawn.

To prove the unbiasedness of tsymm, we first prove that 1 i.e. for any unit i in the population, taken over all samples having unit i drawn first.

(i) n = 2, the probability of jth as second unit when ith unit is drawn in the first draw is





(ii) n = 3, the probability of jth and kth as second and third unit when ith unit is drawn in the first draw.







Similarly for n = 4



Now

 **(7.12.10)**

When we use ΣP(s)tsymm over all samples of size n, the coefficient yi in the sum is 1, hence (7.12.10) becomes

 **(7.12.11)**

For simplicity, we shall consider tsymm from n = 2, then from (7.12.9)

 **(7.12.12)**

where







Substituting these in (7.12.12), we have

 **(7.12.13)**

Alternatively, tsymm for n = 2 may be derived in a simple way. For instance if a sample of two units i and j is selected with PPS without replacement, then units may be selected in two ways.

(i) ith unit first and jth second with 

(ii) jth unit first and ith second with 

and these will be two ordered estimators i.e. tmean (i, j), tmean (j, i).

The unordered estimator tsymm is given by

 **(7.12.14)**

On simplification it comes to be (7.12.13).

***THEOREM (7.8)***

The variance of tsymm for n = 2 is

 **(7.12.15)**

# PROOF

We know that



We have



and 

Now

=  as E(tsymm) = Y

 = 

 = 

 = 

 = 

 

 = 

 

 =

 

 =

 

 = 

 

 = 

 

 = 

 

 = 

 

 = 

 = 

= 

Pathak (1967a) derived the following variance formal for tsymm for any n ≥ 2.

 **(7.12.16)**

Comparing Var(tsymm) and  for n = 2, obviously Var(symm) < .

An unbiased estimator of (7.12.15) is for n = 2.

 **(7.12.17)**

Pathak (1967b) derived the following unbiased variance estimator for any n

**(7.12.18)**

where p(s|ij) denotes the conditional probability of selecting the observed sample s, given that units i, j were selected in that order at the first two draws. This variance formula is no-negative but the computation becomes Cornber some as n increases Bayless (1968) developed a computer programme to calculate p(s|ij), p(s|i) and p(s).

Murthy (1957) further showed that an unordered and therefore more efficient unbiased variance estimator for tmean for n = 2 is

 **(7.12.19)**

Example 7

From the data given in example, draw all possible samples of size 2 under Murthy’s scheme and show that E(tsymm) = Y. Calculate the variance of tsymm.

# SOLUTION

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **All Possible Samples** | **pi** | **pj** | **tmean** | **tsymm** | **tsymm P(s)** |  |
| 1 | 2 | .1 | .2 | 5.45 | 5.4497 | 0.2573 | 1.4025 |
| 1 | 3 | .1 | .3 | 5.90 | 6.1250 | 0.4667 | 2.8585 |
| 1 | 4 | .1 | .4 | 6.35 | 6.8000 | 0.7556 | 5.1381 |
| 2 | 1 |  |  | 5.60 | - | - | - |
| 2 | 3 |  |  | 6.40 | 6.5333 | 1.0500 | 6.8600 |
| 2 | 4 |  |  | 6.80 | 7.1429 | 1.6667 | 11.9051 |
| 3 | 1 |  |  | 6.30 | - | - | - |
| 3 | 2 |  |  | 6.65 | - | - | - |
| 3 | 4 |  |  | 7.35 | 7.5385 | 2.8000 | 21.1078 |
| 4 | 1 |  |  | 7.10 | - | - | - |
| 4 | 2 |  |  | 7.40 | - | - | - |
| 4 | 3 |  |  | 7.70 | - | - | - |
|  |  |  |  |  |  | 7.0000 | 49.3132 |

where



where



Now

E(tsymm) = Σ tsymm P(s) = 7



**(c) Rao, Hartley and Cochran Estimator**

 **[Rao, Hartley and Cochran (1962)]**

In this selection procedure, if sample of size n is to select, the population is divided into n random groups, having Ni (i = 1,2,3,….,n) units in each group, where number of units in each group is predetermined (if possible make all groups equal. One unit is selected from each group. The probabilities of selection being the normal measures of size within the group.

An unbiased estimator of population of Y suggested by Rao, Hartley and Cochran is

 **(7.12.20)**

where piT is the initial probability of the sample unit selected from the n random groups, being the initial probability of the Tth unit of the ith group such that  Since ΣPiT will not be equal for all the groups, the probabilities of selection is not proportional to size.

The derivation of variance of formula of  is very simple. In this selection procedure randomization is involved at two stages, firstly when groups are formed, secondly at stage of selected a sample unit. We can use the well known formula given in Chapter 1 i.e.



Since = Y, which is constant and will have the zero variance of this term, we left with

 **(7.12.21)**

We know that within each group one unit is selected with probability proportional to size, the concept of sampling with probability proportional with replacement can be used, hence the variance formula will

 **(7.12.22)**

 **(7.12.23)**

The probability that two specified units falls in a group of size Ni is Ni(Ni-1)/ N(N-1).

 **(7.12.24)**

using (7.12.21)

 **(7.12.25)**



(7.12.25) may be written as:



 **(7.12.26)**



If N = nR + K where 0 < K < n and R is a positive integer on should choice is N1 = N2 = ….. = Nk < R + 1 and nk+1 = Nk+2 ….. Nn = R [that is to make k groups of size R + 1 and the remaining n – k of size R]. In this case (7.12.26) reduces to

 **(7.12.27)**

If N is a multiple n and K = 0 the variance further reduces to

 **(7.12.28)**

Unbiased estimators of (7.12.27) and (7.12.28) are respectively

 **(7.12.29)**

If Ni = N this reduced

 **(7.12.30)**

Note that  reduces to zero if yi ∝ Pi for all i.

Note that the ratio of  which is the sample with simple random sampling without replacement to with replacement.

This selection procedure is simple and applicable for any sample size.

# Example 8

From the data given in Example 1, find the variance of  for n =2

# SOLUTION

In this example N = 4, n = 2 and



We know that 

Hence



This method suffers slightly in precision but it has its own advantage i.e. simplicity and applicability for any n.

# Example 9

A population of agriculture holders of ten villages for 1967 and 1971 is given. Select a sample or with n = 2, under RHC sampling scheme and estimate total number of agriculture holders for 1971. Find the variance of this estimator.

|  |  |  |  |
| --- | --- | --- | --- |
| **Village** | **1971(Yi)** | **1967(Zi)** | **Cumulated(Zi)** |
| 1 | 60 | 56 | 56 |
| 2 | 55 | 50 | 106 |
| 3 | 60 | 45 | 151 |
| 4 | 70 | 60 | 211 |
| 5 | 75 | 62 | 273 |
| 6 | 65 | 65 | 338 |
| 7 | 50 | 51 | 389 |
| 8 | 60 | 55 | 444 |
| 9 | 65 | 53 | 497 |
| 10 | 80 | 70 | 567 |
|  | 640 |  |  |

# SOLUTION

Since n = 2, the population given under RHC scheme will be divided at random into two groups with random number tables as

|  |  |
| --- | --- |
| **GROUP 1** | **GROUP 2** |
| **Sr.****No.** | **1971****Yi1** | **1967****Zi1** | **Cumulated****Z­i1** |  | **Sr.****No.** | **1971****Yi2** | **1967****Zi2** | **Cumulated****Z­i2** |  |
| 5 | 75 | 62 | 62 | .1093 | 1 | 60 | 56 | 56 | .0988 |
| 6 | 65 | 65 | 127 | .1146 | 2 | 55 | 50 | 106 | .0882 |
| 9 | 65 | 53 | 170 | .0935 | 4 | 70 | 60 | 166 | .1058 |
| 3 | 60 | 45 | 225 | .0794 | 7 | 50 | 51 | 217 | .0899 |
| 10 | 80 | 70 | 295 | .1235 | 8 | 60 | 55 | 272 | .0970 |
|  |  |  |  | π1=.5203 |  |  |  |  | π2=.4797 |

From group 1, one unit is selected between 1 – 295; 4th unit with initial probability, pIT = 45/567 = 0.0794 and second unit from second group between 1 – 272; 3rd unit with probability p2T = 60/567 – 0.1058. Now











Variance of the RHC estimate can be compared with systematic sampling with random arrangement.

If N is a multiple of n (7.12.28) may be written as;

 **(7.12.32)**

and variance of  (derived by Hartley and Rao, 1962) is

 **(7.12.13)**

There is no direct way to compare these estimators. The linear model that used in Section (7.9) will be applied with α = 0.

In order to simply this comparison let N be very large, so that

Yi = βZi **(7.12.33)**

then

 **(7.12.34)**

where C is constant

Now the RHC estimator will be more efficient, equally or less efficient than the Harvitz-Thompson estimator for an exact procedure according as







Now



so that condition that RHC estimator is more efficient is virtually the same as the condition that

 **(7.12.35)**

i.e. 

***Example***

**Actual Data**

|  |  |  |
| --- | --- | --- |
| **Sr.****No.** | **Number of****Schools** | **Enrolment****(000)** |
|  | 38 | 6 |
|  | 85 | 12 |
|  | 43 | 5 |
|  | 56 | 9 |
|  | 22 | 2 |
|  | 38 | 6 |
|  | 51 | 10 |
|  | 53 | 9 |
|  | 36 | 6 |
|  | 64 | 14 |
|  | 40 | 6 |
|  | 99 | 24 |
|  | 54 | 10 |
|  | 53 | 11 |
|  | 38 | 4 |
|  | 16 | 1 |
|  | 7 | 1 |
|  | 18 | 4 |
|  | 51 | 11 |
|  | 30 | 6 |
|  | 46 | 12 |
|  | 64 | 13 |
|  | 27 | 4 |
|  | 23 | 2 |
|  | 9 | 1 |
|  | 111 | 69 |
|  | 105 | 28 |
|  | 83 | 21 |
|  | 83 | 20 |
|  | 58 | 8 |
|  | 94 | 21 |
|  | 50 | 9 |
|  | 33 | 9 |
|  | 28 | 7 |
|  | 74 | 21 |
|  | 132 | 59 |
|  | 39 | 11 |
|  | 50 | 12 |
|  | 17 | 4 |
|  | 49 | 19 |
|  | 109 | 27 |
|  | 108 | 29 |
|  | 117 | 33 |
| **Sr.****No.** | **Number of****Schools** | **Enrolment****(000)** |
|  | 53 | 14 |
|  | 55 | 17 |
|  | 71 | 26 |
|  | 52 | 17 |
|  | 130 | 39 |
|  | 98 | 25 |
|  | 92 | 24 |
|  | 223 | 93 |
|  | 156 | 56 |
|  | 50 | 8 |
|  | 80 | 21 |
|  | 78 | 23 |
|  | 56 | 10 |
|  | 149 | 41 |
|  | 44 | 12 |
|  | 55 | 14 |
|  | 56 | 11 |
|  | 56 | 10 |
|  | 154 | 50 |
|  | 34 | 6 |
|  | 26 | 11 |
|  | 31 | 7 |
|  | 64 | 15 |
|  | 63 | 18 |
|  | 54 | 12 |
|  | 27 | 7 |
|  | 49 | 12 |
|  | 65 | 18 |
|  | 51 | 5 |
|  | 16 | 4 |
|  | 18 | 5 |
|  | 106 | 26 |
|  | 43 | 12 |
|  | 67 | 23 |
|  | 22 | 6 |
|  | 24 | 5 |
|  | 97 | 23 |
|  | 59 | 14 |
|  | 19 | 4 |
|  | 28 | 8 |
|  | 164 | 42 |
|  | 37 | 8 |
|  | 32 | 8 |
| **Sr.****No.** | **Number of****Schools** | **Enrolment****(000)** |
|  | 19 | 3 |
|  | 11 | 2 |
|  | 60 | 14 |
|  | 16 | 3 |
|  | 37 | 7 |
|  | 91 | 19 |
|  | 64 | 16 |
|  | 129 | 33 |
|  | 37 | 9 |
|  | 115 | 28 |
|  | 27 | 3 |
|  | 18 | 2 |
|  | 37 | 2 |
|  | 65 | 10 |

***solution***

**RANDOMLY ARRANGED DATA**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr.****No.** | **No. of****Schools** | **Enrolment****(000)****(Yi)** | **Pi** |
|  | 46 | 12 | 0.00762 |
|  | 43 | 5 | 0.00713 |
|  | 83 | 21 | 0.01376 |
|  | 51 | 5 | 0.00845 |
|  | 132 | 59 | 0.02188 |
|  | 56 | 10 | 0.00928 |
|  | 164 | 42 | 0.02718 |
|  | 31 | 7 | 0.00514 |
|  | 63 | 18 | 0.01044 |
|  | 92 | 24 | 0.01525 |
|  | 27 | 3 | 0.00447 |
|  | 54 | 12 | 0.00895 |
|  | 55 | 14 | 0.00912 |
|  | 50 | 8 | 0.00829 |
|  | 49 | 12 | 0.00812 |
|  | 38 | 4 | 0.00630 |
|  | 80 | 21 | 0.01326 |
|  | 51 | 11 | 0.00845 |
|  | 34 | 6 | 0.00563 |
|  | 24 | 5 | 0.00398 |
|  | 130 | 39 | 0.02154 |
|  | 56 | 11 | 0.00928 |
|  | 33 | 9 | 0.00547 |
|  | 54 | 10 | 0.00895 |
|  | 19 | 3 | 0.00315 |
|  | 37 | 2 | 0.00613 |
|  | 51 | 10 | 0.00845 |
|  | 16 | 3 | 0.00265 |
|  | 7 | 1 | 0.00116 |
|  | 111 | 69 | 0.01840 |
|  | 86 | 20 | 0.01425 |
|  | 38 | 6 | 0.00630 |
|  | 64 | 15 | 0.01061 |
|  | 97 | 23 | 0.01608 |
|  | 37 | 7 | 0.00613 |
|  | 53 | 14 | 0.00878 |
|  | 154 | 50 | 0.02552 |
|  | 71 | 26 | 0.01177 |
|  | 117 | 33 | 0.01939 |
|  | 55 | 17 | 0.00912 |
| **Sr.****No.** | **No. of****Schools** | **Enrolment****(000)****(Yi)** | **Pi** |
|  | 91 | 19 | 0.01508 |
|  | 65 | 10 | 0.01077 |
|  | 94 | 21 | 0.01558 |
|  | 52 | 17 | 0.00862 |
|  | 156 | 56 | 0.02585 |
|  | 26 | 11 | 0.00431 |
|  | 30 | 6 | 0.00497 |
|  | 115 | 28 | 0.01906 |
|  | 67 | 23 | 0.01110 |
|  | 129 | 33 | 0.02138 |
|  | 59 | 14 | 0.00978 |
|  | 23 | 2 | 0.00381 |
|  | 53 | 9 | 0.00878 |
|  | 27 | 7 | 0.00447 |
|  | 149 | 41 | 0.02469 |
|  | 99 | 24 | 0.01641 |
|  | 106 | 26 | 0.01757 |
|  | 56 | 9 | 0.00928 |
|  | 28 | 8 | 0.00464 |
|  | 16 | 1 | 0.00265 |
|  | 37 | 9 | 0.00613 |
|  | 78 | 23 | 0.01293 |
|  | 64 | 13 | 0.1061 |
|  | 44 | 12 | 0.00729 |
|  | 28 | 7 | 0.00464 |
|  | 18 | 4 | 0.00298 |
|  | 64 | 14 | 0.01061 |
|  | 22 | 2 | 0.00365 |
|  | 223 | 93 | 0.00696 |
|  | 108 | 29 | 0.01790 |
|  | 58 | 8 | 0.00961 |
|  | 105 | 28 | 0.01740 |
|  | 40 | 6 | 0.00663 |
|  | 65 | 18 | 0.01077 |
|  | 17 | 4 | 0.00282 |
|  | 18 | 2 | 0.00298 |
|  | 53 | 11 | 0.00878 |
|  | 11 | 2 | 0.00182 |
|  | 19 | 4 | 0.00315 |
|  | 49 | 19 | 0.00812 |
| **Sr.****No.** | **No. of****Schools** | **Enrolment****(000)****(Yi)** | **Pi** |
|  | 39 | 11 | 0.00646 |
|  | 22 | 6 | 0.00365 |
|  | 50 | 9 | 0.00829 |
|  | 16 | 4 | 0.00265 |
|  | 37 | 8 | 0.00613 |
|  | 36 | 6 | 0.00597 |
|  | 38 | 6 | 0.00630 |
|  | 32 | 8 | 0.00530 |
|  | 74 | 21 | 0.01226 |
|  | 18 | 5 | 0.00298 |
|  | 56 | 10 | 0.00928 |
|  | 85 | 12 | 0.01409 |
|  | 9 | 1 | 0.00149 |
|  | 50 | 12 | 0.00829 |
|  | 109 | 27 | 0.01806 |
|  | 98 | 25 | 0.01624 |
|  | 43 | 12 | 0.00713 |
|  | 27 | 4 | 0.00447 |
|  | 60 | 14 | 0.00994 |
|  | 64 | 16 | 0.01061 |

|  |
| --- |
| **Group Number 1** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 46 | 12 | 0.00762 |
|  | 43 | 5 | 0.00713 |
|  | 83 | 21 | 0.01376 |
|  | 51 | 5 | 0.00845 |
|  | 132 | 59 | 0.02188 |
|  | 56 | 10 | 0.00928 |
|  | 164 | 42 | 0.02718 |
|  | 31 | 7 | 0.00514 |
|  | 63 | 18 | 0.01044 |
|  | 92 | 24 | 0.01525 |

|  |
| --- |
| **Group Number 2** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 27 | 3 | 0.00447 |
|  | 54 | 12 | 0.00895 |
|  | 55 | 14 | 0.00912 |
|  | 50 | 8 | 0.00829 |
|  | 49 | 12 | 0.00812 |
|  | 38 | 4 | 0.00630 |
|  | 80 | 21 | 0.01326 |
|  | 51 | 11 | 0.00845 |
|  | 34 | 6 | 0.00563 |
|  | 24 | 5 | 0.00398 |

|  |
| --- |
| **Group Number 3** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 130 | 39 | 0.02154 |
|  | 56 | 11 | 0.00928 |
|  | 33 | 9 | 0.00547 |
|  | 54 | 10 | 0.00895 |
|  | 19 | 3 | 0.00315 |
|  | 37 | 2 | 0.00613 |
|  | 51 | 10 | 0.00845 |
|  | 16 | 3 | 0.00265 |
|  | 7 | 1 | 0.00116 |
|  | 111 | 69 | 0.01840 |

|  |
| --- |
| **Group Number 4** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 86 | 20 | 0.01425 |
|  | 38 | 6 | 0.00630 |
|  | 64 | 15 | 0.01061 |
|  | 97 | 23 | 0.01608 |
|  | 37 | 7 | 0.00613 |
|  | 53 | 14 | 0.00878 |
|  | 154 | 50 | 0.02552 |
|  | 71 | 26 | 0.01177 |
|  | 117 | 33 | 0.01939 |
|  | 55 | 17 | 0.00912 |

|  |
| --- |
| **Group Number 5** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 91 | 19 | 0.01508 |
|  | 65 | 10 | 0.01077 |
|  | 94 | 21 | 0.01558 |
|  | 52 | 17 | 0.00862 |
|  | 156 | 56 | 0.02585 |
|  | 26 | 11 | 0.00431 |
|  | 30 | 6 | 0.00497 |
|  | 115 | 28 | 0.01906 |
|  | 67 | 23 | 0.01110 |
|  | 129 | 33 | 0.02138 |

|  |
| --- |
| **Group Number 6** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 59 | 14 | 0.00978 |
|  | 23 | 2 | 0.00381 |
|  | 53 | 9 | 0.00878 |
|  | 27 | 7 | 0.00447 |
|  | 149 | 41 | 0.02469 |
|  | 99 | 24 | 0.01641 |
|  | 106 | 26 | 0.01757 |
|  | 56 | 9 | 0.00928 |
|  | 28 | 8 | 0.00464 |
|  | 16 | 1 | 0.00265 |

|  |
| --- |
| **Group Number 7** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 37 | 9 | 0.00613 |
|  | 78 | 23 | 0.01293 |
|  | 64 | 13 | 0.01061 |
|  | 44 | 12 | 0.00729 |
|  | 28 | 7 | 0.00464 |
|  | 18 | 4 | 0.00298 |
|  | 64 | 14 | 0.01061 |
|  | 22 | 2 | 0.00365 |
|  | 223 | 93 | 0.03696 |
|  | 108 | 29 | 0.01790 |

|  |
| --- |
| **Group Number 8** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 58 | 8 | 0.00961 |
|  | 105 | 28 | 0.01740 |
|  | 40 | 6 | 0.00663 |
|  | 65 | 18 | 0.01077 |
|  | 17 | 4 | 0.00282 |
|  | 18 | 2 | 0.00298 |
|  | 53 | 11 | 0.00878 |
|  | 11 | 2 | 0.00182 |
|  | 19 | 4 | 0.00315 |
|  | 49 | 19 | 0.00812 |

|  |
| --- |
| **Group Number 9** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 39 | 11 | 0.00646 |
|  | 22 | 6 | 0.00365 |
|  | 50 | 9 | 0.00829 |
|  | 16 | 4 | 0.00265 |
|  | 37 | 8 | 0.00613 |
|  | 36 | 6 | 0.00597 |
|  | 38 | 6 | 0.00630 |
|  | 32 | 8 | 0.00530 |
|  | 74 | 21 | 0.01226 |
|  | 18 | 5 | 0.00298 |

|  |
| --- |
| **Group Number 10** |
| **Sr.****No.** | **Zi** | **Yi** | **Pi** |
|  | 56 | 10 | 0.00928 |
|  | 85 | 12 | 0.01409 |
|  | 9 | 1 | 0.00149 |
|  | 50 | 12 | 0.00829 |
|  | 109 | 27 | 0.01806 |
|  | 98 | 25 | 0.01624 |
|  | 43 | 12 | 0.00713 |
|  | 27 | 4 | 0.00447 |
|  | 60 | 14 | 0.00994 |
|  | 64 | 16 | 0.01061 |

|  |
| --- |
| **Necessary Calculations for Estimating****Total No. of Students with its Variance** |
| **Selected****Units** | **Enrolment****yi (000)** | **Pi** | **π** | **yit/Pit \*π** | **(yit/Pit-y’)2** | **π(yit/Pit-y’)2** |
| 1 | 12 | 0.00762 | 0.12612 | 198.5217 | 11571.8549 | 1459.426846 |
| 13 | 14 | 0.00912 | 0.07657 | 117.6000 | 4818.1469 | 368.9068392 |
| 28 | 3 | 0.00265 | 0.08518 | 96.3750 | 112318.417 | 9567.727305 |
| 33 | 15 | 0.01061 | 0.12794 | 180.9375 | 2734.83500 | 349.8993415 |
| 43 | 21 | 0.01558 | 0.13673 | 184.3085 | 14040.6202 | 1919.706951 |
| 57 | 26 | 0.01757 | 0.10209 | 151.0943 | 182.880612 | 18.66994652 |
| 62 | 23 | 0.01293 | 0.11369 | 202.2821 | 97807.5649 | 11119.65356 |
| 79 | 4 | 0.00315 | 0.07209 | 91.5789 | 38493.8947 | 2775.081907 |
| 88 | 8 | 0.00530 | 0.05999 | 90.5000 | 1762.79060 | 105.7557507 |
| 96 | 25 | 0.01624 | 0.09960 | 153.3163 | 5295.66418 | 527.4600885 |





Estimated Total = 1466.5144 Thousand

Sample Variance = 2821.2289

Standard Error of Estimate = 53.1152

Population Total = 1557 Thousand

Actual Variance of Estimate = 233094.5289

Actual Standard Error = 482.7986

**(d) Lahiri’s Estimator**

 **[Lahiri (1951)]**

In chapter 6, we know that ratio estimator with equal probability sampling selection is biased. Lahiri (1951) given a sampling selection procedure of with probability proportional to the aggregate of the sizes of the sample units. The procedure may be described as a set of n units using simple random sampling without replace merit as selected and size measure of those units are aggregated. A random number between, zero and seem of the sizes of n largest units (or any number greater than this) is chosen. If this random number exceeds the aggregate size of the sample random sampling without replacement of n units, the sample is reject as a whole otherwise accepted of the sample is rejected, the process is repeated until a sample is accepted.

A simple procedure of selecting a sample with probability proportional to the aggregate of the size has been proposed by Midzuno (1952), Sen (1952). In which procedure the unit is selected proportional to size and the remaining (n – 1) units from the (N – 1) with simple random sampling without replacement. The probability will be

 **(7.12.36)**

as the first unit is selected with probability and the remaining (n – 1) units of the sample with probability 

When the selection of a sample is proportional to its aggregate measure of size, the conventional ratio estimator

 **(6.1.2)**

is unbiased as







as each population unit occur  samples. For n = 2 the variance of y”2 is

 **(7.12.37)**

Raj (1954) and Sen (1955) provided an unbiased variance estimator for n = 2 i.e.

 **(7.12.38)**

which can take negative values.

Rao and Vijayan (1977) have proposed his new unbiased estimators, which for some populations are non negative. For the case n = 2 these estimators coincide and take the form

 **(7.12.39)**

where

 **(7.12.40)**

For n > 2 both the estimator are different. For details see Rao and Vijayan (1975).

***Example 10***

Draw all possible distinct samples for n = 2 from the data given in Example 1. Constant Lahiri estimator for each sample and show that and calculate the variance of .

***SOLUTION***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **All Possible** | **pi** | **pj** | **yi** | **yj** |  |  |  |
| 1 | 2 | .1 | .2 | .5 | 1.2 | 5.6667 | 1.7 | 9.6334 |
| 1 | 3 | .1 | .3 | .5 | 2.1 | 6.5000 | 2.6 | 16.9000 |
| 1 | 4 | .1 | .4 | .5 | 3.2 | 7.4000 | 2.7 | 27.3800 |
| 2 | 3 | .2 | .3 | 1.2 | 2.1 | 6.6000 | 3.3 | 21.7800 |
| 2 | 4 | .2 | .4 | 1.2 | 3.2 | 7.3333 | 4.4 | 32.2665 |
| 3 | 4 | .3 | .4 | 2.1 | 3.2 | 7.5714 | 5.3 | 40.1284 |
|  |  |  |  |  |  |  | 21 | 148.08833 |



To prove the  we see that each unit of the population appearing in times in all the sample





= 49.363 – 49 = 0.363

This can be calculated directly using the expression (7.12.37).

**(e) Poisson Sampling**

 **[Hajek (1964)]**

Poisson sampling is a name given to a selection procedure in which all the population units have independent, and in general un-equal, probabilities of selection is πi, Poisson sampling as defined by Hajek (1964) gives each unit in the population a certain probabilities of inclusion in the sample which is denoted by πi for the ith unit. To select, a set of N Bernoulli trials is carried out to determine whether each unit in term is to be included in the sample or not. It was actually used by the U.S. Bureau of Census for its Annual Survey of Manufacturer (Ogus and Clark 1971). Poisson sampling is known in forestry as 3-P sampling (Sethumadhavi and Rajagopalen 1974). For the Bureau, the chief virtue of Poisson sampling lay in the simple manner in which control could be exercised over the decision as to which units were to be included in various samples. This property of Poisson sampling has been considered in detail by Brewer, Early and Joyee (1972) and Sunter (1977). Mathematical treatment was given by Brewer, Early and Hanif (1980). Poisson sampling suffers from the draw back that the sample size m is a random variable.

The unbiased Horvitz-Thompson estimator of the population total Y is

 **(7.12.36)**

where s is the set of units in the sample.

Since the joint probability of inclusion πij takes the simple form πij = πiπj (because of independent), the variance of (7.12.36) may be derived easily, using (7.8.1) and putting πij = πiπj i.e.

 **(7.12.37)**

An unbiased estimator of (7.12.37) is from (7.8.4)

 **(7.12.38)**

Because the sample size m varies in sampling procedure, the ratio estimator

 if m > 0 where E(m) = n

 . 0 otherwise **(7.12.38)**

is more efficient for large samples. If Yi are roughly proportional to πi, it is more efficient for samples of any size.

It can be show that for large n the mean square error of is given

 **(7.12.40)**

where Po = Pr (m = o)

The conventional estimator of (7.12.40) is

 **(7.12.41)**

Although the probability of selecting an empty sample is trivially small when n is large, the problem cannot be ignored in all large scale surveys. This is because such surveys typically employ a fairly detailed stratification by type of unit, and the target sample size within some of the smaller strata is often quite modest. The modified form has been suggested by Ogus and Clark (1971) which ensures that an empty sample is never selected. The name given by them MODIFIED POISSON SAMPLING. In this an ordinary Poisson sample is drawn first, but if there are 200 units in that sample, a second Poisson sample is drawn and so on repeatedly until a non empty sample is achieved.

The only advantage of modified Poisson sampling over ordinary Poisson sampling is that it ensures a non-empty. If the sample selected is much smaller (or much larger) than the target size, modified Poisson sampling provides no remedy. A procedure which ensures a much more stable sample size is called COLLOCATED SAMPLING. This is similar to Poisson sampling but reduces the variation in sampling size by requiring the random variable ri to be uniformly spaced instead uniformly distributed over the interval (0, 1). A random number L (Li = 1, 2, 3, …. N) is chosen with equal probabilities, and a random variable γ is also selected from a uniform distribution over the interval (0, 1). For each i we then define.

 **(7.12.42)**

The Horvitz-Thompson estimator still used, but now no simplification of its variance formula is possible. (For details see Brewer, Early and Hanif 1980).

**7.13 A GENERALIZATION OF THE HORVITZ-THOMPSON ESTIMATOR**

Uptil now in this Chapter, unequal probabilities with and without replacement has been considered. But some selection procedures cannot be categorized as either with replacement or without replacement in the usual sense. The most important there are intermediate cases where for example, one or more of the population units may appear more than once in sample but the remaining units appear at most once. Hanif and Brewer (1979) presented a general theory of sampling with unequal probabilities which allows population units to appear more than once in sample. Examples of these (i) ordinary optimatic selection where one or more of the population units is large enough to be certain of selection at least once (ii) Deming’s (1960) procedure which selects several systematic samples with different random starts.

Let S­i be the number 9 times the ith population unit appear in sample and Sij the number of times the ordered pair (i, j) appears in the set of n(n-1) ordered pairs of sample units. Then



The expected values of Si and Sij will be written as μi and μij respectively. Generalized Horvitz-Thompson (GHT) estimated may be defined as

 (**7.12.42)**

which is clearly unbiased as



The variance of the GHT estimator is







[E(Si)2 = μii + μi and μij = 0]

 **(7.13.2)**

which may be written as

 **(7.13.3)**

This is similar in form to (7.8.3) but general to its meaning. When selection is strictly without replacement μi = πi, μij = πij for j ≠ i and μij = 0; then (7.13.3) is identical with (7.8.1). Waiting for convenience Pi = μi/n, Pij = μij/n(n – 1) (7.13.2) or (7.13.3) may be written as

 **(7.13.4)**

and

 **(7.13.5)**

For sampling with replacement μij = n(n – 1) PiPj (7.13.4) reduces to (7.4.1) which is variance formula for unequal probabilities with replacement.

Many optimal properties possessed by the Horvitz-Thompson estimator are not carried over to GHT. The Hansen-Hurvitz estimator, though convenient and widely used, is well known to be inadmissible, and this will generally be true of an estimator for which the Si­ can take values other than 0 and 1.

The generalization of the Sen-Yates-Grundy variance estimator is

 **(7.13.6)**

**7.14 EMPIRICAL AND SEMI EMPIRICAL COMPARISON**

 **OF VARIOUS ESTIMATORS**

We have mentioned several estimators in this chapter. Since it is difficult to compare these estimators theoretically, we present some empirical and semi empirical comparison. To start with two examples (Yates and Grundy, 1953) are presented along with variances of different estimators.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Zi | 0.1 | 0.2 | 0.3 | 0.4 |
| Population 1 | Y­i | 0.8 | 1.4 | 1.8 | 2.0 |
| Population 2 | Y­i | 0.2 | 0.6 | 0.9 | 0.8 |
| Following are the variances of different estimators under various selection procedures for both the populations. |
| Selection Procedures | Population 1 | Population 2 |
| Hansen-Hurvitz Estimator |  |  |
| P.P.S. with replacement | 500 | 0.125 |
| Hurvitz-Thompson Estimator |  |  |
| Yates and Grundy | 0.057 | 0.059 |
| B.D.S. | 0.275 | 0.058 |
| Random Systematic | 0.367 | 0.033 |
| Midzuno | 0.384 | 0.240 |
| Raj’s estimator | 0.365 | 0.088 |
| Murthy’s estimator | 0.312 | 0.070 |
| RHC estimator | 0.333 | 0.083 |
| Lahiri estimator | 0.510 | 0.101 |

From this empirical studies with N = 4, n = 2, Pi = 0.1, 0.2, 0.3, 0.4, it is different rather not far to draw any definite conclusion, which estimator is more efficient.

Rao and Bayless (1969) and Bayless and Rao (1970) conducted both empirical and semi empirical studies using the linear model (7.9.1) with α = 0. They found that there were not applicable differences of the efficiency of the Harvitz and Thompson estimator in practice from one selection procedure to another. Brewer and Hanif (1969a) for n = 2, reached the same conclusion.

Hanurav (1962), Vijayan (1966) compared the relative efficiencies under the model (7.9.1) ignoring α of and tsymm and came to the conclusion that  was better than tsymm for γ = 0.5. Rao (1966) further proved that  was better than  for all values of γ. Rao and Barles (1969) and Bayless and Rao (1970) in their semi empirical studies, they found that t­symm­ was nearly always more efficient than the . They used the value of γ = 0.5, 0.75, 0.85 and 1. For all values of γ and for n = 2, they found that Murthy’s estimator was consistently more efficient than Raj’s estimator. For γ > 0.5. Murthy’s estimator was more efficient than for γ < 0.875 and less efficient for γ = 1. For n = 3, and 4, Raj’s estimator was less efficient than  for all values of γ. Murthy’s estimator was again more efficient than the  for γ < 0.875. Samiuddin, Hanif and Asad (1978) studied the behaviour of tsymm,  and several other estimator with six artificial and six semi empirical populations. They found that Harvitz-Thompson estimator was reasonably more efficient in almost all cases.

In the semi-empirical studies carried out by the Rao and Bayless (1969), the was found to be consistently less efficient than both tsymm and , its efficiency was greatly affected by the small value of γ. As with Murthy’s and Raj’s estimators, most of the differences were only of the order of a few percent. They also concluded that Murthy’s variance estimator was consistently more stable than the Sen-Yates-Grundy variance estimator for all value of γ. Murthy’s variance estimator also tended to be more stable than Raj’s variance estimator especially for large value of γ and of n. The RHC variance estimator was consistently more stable than Raj, Murthy and Sen-Yates-Grundy variance estimator for all values of γ, however the gains over Murthy’s estimator was not large. For n = 3 and 4 RHC variances estimator was still almost always more stable than Murthy’s variance estimator for γ = 0.875 but for γ = 1 the case was reverse. They found that the Lahiri’s estimator was more efficient than the , t­mean, tsymm and when either (i) few units in the population had large sizes relative to the sizes of remaining units in the population, and sample containing those units gave good estimates of Y, or (ii) the coefficient of variation of the benchmark variable was small Bayless and Rao (1970) extended their investigation for n = 3 and 4, both with respect to the efficiency of the estimator of total and poor performance of the variance estimator.