*CHAPTER 7*

**SAMPLING WITH PROBABILITIES PROPORTIONAL TO SIZE (WITH REPLACEMENT)**

**7.1. INTRODUCTION.**

In previous chapters equal probability sampling selection procedure and estimation methods have been discussed. In this and subsequent chapters those selection procedures will be considered in which probability of selection varies from unit to unit (unequal probability) in the population. In equal probability sampling, selection does not depend how large or small that unit is but in probability proportionate (proportional) to size sampling it does.

The 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* **(** i.e. **)** is significantly correlated with *estimand,*( variable under study) Yi. A method of selection in which the units are selected with probability proportionate (proportional) 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 use of unequal probabilities in sampling was first suggested by Hansen and Hurwitz (1943). Prior to that there had been substantial developments in sampling theory and practice, but all these had been based on the assumption that the probabilities of selection within each stratum are equal. Hansen and Hurwitz (1943) proposed a two stage sampling scheme (will be discussed in Chapter 11). The first stage selection took place in independent draws. At each draw, a single first-stage unit is selected with probabilities proportional to a given measure of 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 units. Because it is possible for the same first-stage unit to be selected more than once therefore, 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 sampling can however be used in single stage design.

The selection of sample with unequal probabilities is explained below:

Population of males and females of 523 villages of Multan district along with area of each village is given in Appendix-I. In order to understand the selection procedure of probability proportional to size sampling, 5% sample has been selected from this population. For the selection of the sample we cumulate the measure of sizes (area), 26(5% of total villages) random numbers are selected from 001 to 956204(total of area). These random numbers along with the serial number of villages, total population and initial probabilities of selection are given (data is given on next page). If any unit is selected more than once it should be included in the sample

* 1. **THE HANSEN – HURWITZ ESTIMATOR.**

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

 **(7.3.1)**

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

We now prove the unbiasedness and derive the variance of (7.3.1) in the following theorem.

**THEOREM (7.1)**

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

**PROOF**

We know that

 **(7.3.1)**

Taking the expectation

◊

Therefore  is an unbiased estimator of population total Y.

**Table 7.1: PPS Sample of Villages**

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

**7.4. VARIANCE AND UNBIASED VARIANCE ESTIMATOR**

**THEOREM (7.2)**

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



Substituting the value of  from (7.3.1), we have





.

Since the selection of population units are independent; therefore Pij = PiPj, substituting the value of :

.

On simplification we get:

◊

This expression may alternatively be written as

. (7.4.2)

 (7.4.3)

. (7.4.4)

* + 1. **An Alternative Proof( Using Indicator Variable)**

Expression(7.4.2) may also be derived by using the indicator variables:

Let ai be the number of times the ith unit of the population is in the sample (Chapter 2), then the joint distribution of ai is given as:

 (7.4.5)

Further

 (7.4.6)

Using indicator variable the estimator (7.3.1) may be written as:

 (7.4.7)

The unbiasedness can be proved easily by taking the expectation of (7.4.7) and putting E(ai) = nPi from (7.4.6); that is:

◊

The variance of  may be written (see Chapter 2) as:

(7.4.8)

Putting the values of  and  from (7.4.6) in (7.4.8) and on simplification we get (7.4.1). It follows that, if  the variance of  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.

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

. (7.4.9)

**7.4.2 Unbiased Variance Estimator**

**THEOREM (7.3)**

A sample of size n is drawn from a population of N units with probabilities proportional 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)

,

Now

.

Applying expectation we have:





Using (7.4.2) we get

.

Using this result in (7.4.10), we get

◊

Expression (7.4.10) may also 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 analogous to (7.4.9) as

. (7.4.13)

Though the Hansen–Hurwitz (1943) scheme is based on sampling with replacement process but for the following reasons, it is preferred to be used in large scale sample surveys;

1. the selection of the sample is simple,
2. it 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 estimated i.e. Yi and Zi are linearly related and regression line passes through origin.

**EXAMPLE (7.1)**

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

|  |  |  |  |
| --- | --- | --- | --- |
| no | Yi | Zi |  |
| 1 | 0.5 | 1 | 0.1 |
| 2 | 1.2 | 2 | 0.2 |
| 3 | 2.1 | 3 | 0.3 |
| 4 | 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 |

Now 

And  = 49.50 – 49 = 0.50

Also by using (7.4.1) we have:



**EXAMPLE 7.2:**

Following is a sample of 26 villages selected by using probability proportional to size sampling and with replacement selection procedure form the data given in Appendix-I. Estimate the total number of person in 523 villages and compare this result with actual number of population given in 523 villages also calculate standard error of this estimate:

***Solution***: Necessary calculations are given in the table below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sr. No. |  | =z/Z |  |  |  |  |
| 1 | 7346 | 0.005946 | 1235452.405 | 137886694865.41 | 1.52634E+12 | 9075633367 |
| 2 | 9231 | 0.006511 | 1417754.569 | 35731892260.18 | 2.01003E+12 | 13087292428 |
| 3 | 3713 | 0.001335 | 2781273.408 | 1379426820568.15 | 7.73548E+12 | 10326868165 |
| 4 | 2310 | 0.001337 | 1727748.691 | 14632605885.20 | 2.98512E+12 | 3991099476 |
| 5 | 7261 | 0.006127 | 1185082.422 | 177831700028.54 | 1.40442E+12 | 8604883467 |
| 6 | 10425 | 0.00353 | 2953257.79 | 1812993331051.22 | 8.72173E+12 | 30787712465 |
| 7 | 6978 | 0.006409 | 1088781.401 | 268326052675.04 | 1.18544E+12 | 7597516617 |
| 8 | 399 | 0.000316 | 1262658.228 | 118422122062.40 | 1.59431E+12 | 503800632.9 |
| 9 | 737 | 0.002414 | 305302.403 | 1693852740901.42 | 93209557065 | 225007870.8 |
| 10 | 3203 | 0.001396 | 2294412.607 | 472833951008.72 | 5.26433E+12 | 7349003582 |
| 11 | 4039 | 0.002813 | 1435833.63 | 29223818023.56 | 2.06162E+12 | 5799332030 |
| 12 | 5439 | 0.000906 | 6003311.258 | 19329457362517.20 | 3.60397E+13 | 32652009934 |
| 13 | 1373 | 0.000885 | 1551412.429 | 3065942449.29 | 2.40688E+12 | 2130089266 |
| 14 | 8074 | 0.006968 | 1158725.603 | 200755773987.28 | 1.34265E+12 | 9355550517 |
| 15 | 3416 | 0.00166 | 2057831.325 | 203444246707.55 | 4.23467E+12 | 7029551807 |
| 16 | 5841 | 0.003874 | 1507743.934 | 9808812374.83 | 2.27329E+12 | 8806732318 |
| 17 | 1316 | 0.002297 | 572921.202 | 1068871009157.62 | 3.28239E+11 | 753964301.3 |
| 18 | 6475 | 0.004451 | 1454729.274 | 23120451813.51 | 2.11624E+12 | 9419372051 |
| 19 | 1261 | 0.002064 | 610949.612 | 991684897660.22 | 3.73259E+11 | 770407461.2 |
| 20 | 6975 | 0.006409 | 1088313.309 | 268811216688.66 | 1.18443E+12 | 7590985333 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sr. No.** |  |  |  |  |  |  |
| 21 | 2513 | 0.002781 | 903631.787 | 494422166072.03 | 8.1655E+11 | 2270826681 |
| 22 | 3039 | 0.001128 | 2694148.936 | 1182363847314.18 | 7.25844E+12 | 8187518617 |
| 23 | 322 | 0.000697 | 461979.910 | 1310574981674.20 | 2.13425E+11 | 148757532.3 |
| 24 | 13056 | 0.00613 | 2129853.181 | 273602014185.76 | 4.53627E+12 | 27807363132 |
| 25 | 593 | 0.000998 | 594188.377 | 1025348645657.47 | 3.5306E+11 | 352353707.4 |
| 26 | 2515 | 0.001936 | 1299070.248 | 94687373182.90 | 1.68758E+12 | 3267161674 |
|  | **117850** | **0.081318** | **41776367.9425** | **32621180470772.50** | **99746754265786.8** | **217890794431.5760** |

(i) Estimated Total 

****

The actual/total for 523 villages is 1797841.

(ii) 



(iii) 

(iv) 

This may also be calculated as:

= 

=  = 50186431493

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

**and Probability Proportional to Size with Replacement**

The efficiency of probability proportional to size sampling with replacement relative to simple random sampling with replacement can be easily studied as under:

We know that

 (7.4.1)

Substituting Pi =1/N in (7.4.1) we have

 (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 get

, (7.4.15)

where .

We can see from (7.4.15) that probability proportional to size (PPS) sampling with replacement will be more efficient than simple random sampling with replacement if

 (7.4.16)

i.e. If Zi and  are positively correlated.

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

**7.4.3Comparison of** **and** **Using a Linear Stochastic Model:**

In this section we have compared the mean per unit estimator of simple random sampling and the Hansen – Hurwitz estimator under the linear stochastic model.

We have already shown in (7.4.16) that:

 (7.4.15)

 (7.4.17)

For the purpose of comparison, let us take the linear model defined in (6.8.2) as:



 (6.8.2)

Substituting the value of Yi from (6.8.2) in (7.4.17), we have



Taking model expectation



Using we get





 (7.4.18)

From (7.4.18) it is concluded that PPS sampling with replacement is more efficient than simple random sampling, if



or 

This satisfied only if , since σ2 ≥ 0 and 

Or  (7.4.19)

Alternatively above result can be achieved as under:

We know that ,

Summing over i we have, 

Now, the variance for population total for simple random sampling with replacement (ignoring fpc) is

.

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



or 



Since  and , therefore

 (7.4.20)

Now  (7.4.1)

Putting  we get

 (7.4.21)

Putting the values of Yi and Y from the model and taking the expectation under the condition of model we have:



 (7.4.22)

Since ; so (7.4.22) transforms to:

 (7.4.23)

From (7.4.20) and (7.4.23) we have







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



which is same as (7.4.19).

* 1. **GAIN IN PRECESION DUE TO PPS SAMPLING (WITH REPLACEMENT) OVER SIMPLE RANDOM SAMPLING**

The gain in precision of probability proportional to size sampling with replacement over simple random sampling with replacement can be easily obtained as under. We know that:

 (2.5.2)

and  (7.4.12)

It can be easily proved that

(i)  (7.5.1)

and



. (7.5.2)

Using (7.5.1) and (7.5.2) in (2.5.2) we have:

.

(7.5.3)

.

. (7.5.4)

Subtracting  from (7.5.4) we get

.

.



. (7.5.5)

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

. (7.5.6)

**Example (7.3)**

A sample of size 5 has been selected from a population of 20 farms. Number of trees, along with initial probability of selection is given

i) Estimate the total number of trees in the area, calculate the estimated variance and standard error of this estimator.

ii) Estimate the gain in precession over simple random sampling. The actual number of trees is 28443.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No. of** Villages | **No. of Trees (yi)** | **Probability of Selection (pj)** |  |  |  |
| 8  4  16  11  10 | 311  949  11799  2483  3044 | 0.014  0.036  0.275  0.121  0.212 | 22214.286  26361.111  42905.455  20520.661  14358.490 | 9349614.91  1186162.77  310938735.20  22575222.29  119104700.50 | 6908642.9  25016694.4  506241458.1  50952801.6  43707245.28 |
|  |  |  | **126360.003** | **463154435.5** | **632826842.1** |

(i) Estimated Total  .



Actual Total 

(ii) 



(iii) 

 (7.5.4)



= 383158187.5





* 1. **ALTERNATIVE ESTIMATOR OF POPULATION TOTAL**

Pathak (1962) has described an alternate estimator of population total in a probability proportional to size sampling with replacement .This estimator for a sample of size 3 is given as

, (7.6.1)

For sample size n this estimator is:

.. (7.6.2)

This estimator is more efficient than Hansen and Hurwitz (1993) estimator but is more difficult to calculate. The gain in precision is small unless the sampling fraction is large.

* 1. **RATIO ESTIMATION IN PPS SAMPLING**

As in case of simple random sampling the ratio method of estimation brings considerable decrease in error therefore same method can be used in probability proportional to size sampling to increase precision. For this consider:



Then the ratio estimator in PPS sampling can be written as:

. (7.7.1)

From Hansen, Hurwitz and Madow (1953) we have

.

(7.7.2)

Using (7.4.2) and (7.4.9) and analogues expression

, (7.7.3)

in (7.7.2) and on simplification we get:

 (7.7.4)

. (7.7.5)

This may be put easily as

. (7.7.6)

An approximate unbiased estimator of  may be written in a straight forward way as:

, (7.7.7)

or . (7.7.8)

* 1. **LAHIRI’S SELECTION PROCEDURE**

The method of selection of sample in PPS with replacement requires cumulation of measure of size. To avoid this Lahiri(1951) has proposed an alternative method of selection of sample. This method is given as:

Suppose population of N units with measures of sizes Zi is given. Let 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 from 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 with replacement 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  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 is; 

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.

In the following we have given as example to explain this selection procedure.

**EXAMPLE( 7.4)**

Suppose the pair of selected random numbers be (11, 68), the size of the 11th number is less than 68 (table on next page) 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

**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 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. of  Village | No. of Trees  Yi | Area in Acres Zi |  | Effective  Range | In-effective  Range |
|  | 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’ selection procedure 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 is:



Therefore the total number of estimated trees is 28836.

Also 

 = 38478561

So 

**7.9. STRATIFICATION**

We know that stratification is the most frequently used device for improving the “representativeness” of a sample. It amounts to forming the population is to be splitted into two or more groups, known as strata, and selecting a sample from each independently. Intuitively, the virtue of stratification is seen as a purposeful “balancing” of the sample so as to represent each of the strata and to represent them in a desired fashion, rather than (as in simple random sampling) leaving the representation to chance. Intuitively too, such a procedure only has merit where the stratification process results in the strata having significantly different characteristics. From another viewpoint the stratified sampling is seen as the elimination of the “between strata” component of total variance.

In simple random sampling, it necessary to have a complete list of all the elements in the population in addition, if stratification is to be employed to good effect, it is also necessary to have some supplementary information regarding each element in the population. This information may be qualitative or quantitative. Without relevant supplementary information, stratification cannot be advantageous.

Selection with unequal probabilities within each of the strata can often be useful i.e. geographical stratification.

To obtain maximum benefits from stratification, it would be desirable to form as many strata as there are units to be selected in the sample and to draw one sample unit from each. However there are two drawbacks of this procedure. First, it is not possible with only one sample unit per stratum to make sample estimates of variance in the ordinary manner. Secondly, any failure through operational difficulties to obtain results for any sample unit creates serious difficulties in estimation. Consequently it is a frequent practice to form only half as many strata as there are units to be selected, and to select two units from each.

The following formulae for stratified random sample may be written in a usual way using the concept given in chapter 3.



and



Therefore Hansen and Hurwitz unbiased estimator can be written in a straight forward manner using the concept of stratified random sample (chapter 3)

 (7.9.1)

The , for stratified random sampling is:

 (7.9.2)

or  (7.9.3)

or  (7.9.4)

and  (7.9.5)

Unbiased estimator corresponding to (7.9.3) is:

 (7.9.6)

or  (7.9.7)

A ratio estimator, , for population total in stratified random sampling with unequal probabilities is:

 (7.9.8)

The  for probability proportional to size sampling corresponding to (7.9.4) is:

 (7.9.9)

and corresponding to (7.9.5) is

 (7.9.10)

An approximate unbiased estimator for (7.9.10) for stratified random sampling is:

 (7.9.11)

**EXERCISES**

7.1 Following is the population of 14 clusters. The number of farms and number of cattle are recorded. Select a sample of size 3 using

1. P.P.S. sampling with replacement
2. Pathak’s selection procedure
3. Lahiri’s selection procedure

Estimate the total number of cattle in the population with the help of each selection procedure and find the standard error in each case. Also find the average number of cattle per farm.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **Zi** | **Y­i** |
|  | 32 | 351 |
|  | 83 | 906 |
|  | 18 | 316 |
|  | 30 | 287 |
|  | 55 | 914 |
|  | 24 | 284 |
|  | 66 | 598 |
|  | 48 | 359 |
|  | 64 | 784 |
|  | 30 | 393 |
|  | 40 | 489 |
|  | 70 | 516 |
|  | 48 | 793 |
|  | 25 | 401 |
|  | 633 | 7391 |

* 1. Prove that an unbiased estimator of (7.4.1) is



7.3 The following is the data regarding wheat cultivation of 33 villages for the two different years. Select 20% sample from this population using simple random sampling and using P.P.S. sampling; estimate the total area under wheat for the year 1975 and find the standard error in each case. Also find the relative efficiency.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **1975 Acres**  **(Y)** | **1970 Acres**  **(Z)** | **1975**  **(Y)** | **1970**  **(Z)** | **1975**  **(Y)** | **1970**  **(Z)** |
| 501 | 52 | 634 | 149 | 1194 | 289 |
| 1770 | 381 | 1060 | 278 | 827 | 111 |
| 1737 | 634 | 1060 | 278 | 360 | 112 |
| 946 | 355 | 470 | 99 | 1625 | 498 |
| 827 | 111 | 96 | 6 | 1304 | 399 |
| 377 | 79 | 256 | 105 | 186 | 27 |
| 1767 | 515 | 604 | 249 | 701 | 85 |
| 524 | 221 | 715 | 179 | 845 | 330 |
| 571 | 133 | 962 | 144 | 407 | 103 |
| 1016 | 219 | 184 | 62 | 282 | 79 |
| 194 | 60 | 439 | 100 | 854 | 141 |

7.4 Following are the two populations with the same measure of size. A sample of size 2 is taken from each population using P.P.S. sampling with replacement. Find variance for both populations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Measure of size | 1 | 1 | 2 | 3 | 3 |
| Population 1 | .3 | .5 | .8 | .9 | 1.5 |
| Population 11 | .3 | .3 | .8 | 1.5 | 1.5 |

­7.5 The purpose is to estimate the total number of households in an area containing the total number of household blocks. The eye estimated (Z) number of households in each block is obtained by making a rapid survey of the area. The actual Y number of households is later obtained through careful field work.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Block | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Y | 19 | 9 | 17 | 13 | 21 | 22 | 27 | 35 | 20 | 15 | 18 | 37 | 42 | 47 | 27 | 25 | 25 | 25 | 13 | 9 |
| Z | 19 | 9 | 14 | 20 | 24 | 25 | 23 | 24 | 17 | 14 | 18 | 40 | 12 | 30 | 27 | 26 | 21 | 24 | 9 | 19 |

Select a sample of 3 units using Hansen and Hurwitz selection procedure. Estimate the total households and find the standard error of this cabin and compare it with simple random sampling.