***CHAPTER 6***

# RATIO AND REGRESSION ESTIMATORS

**6.1 INTRODUCTION**

In the previous Chapters we have considered the problems associated with the estimation of a single population characteristics. If ancillary or supplementary information is available for each of the units of the population, it can, under suitable conditions, be used in several ways to improve efficiency of the estimators of the variable under investigation (estimand). The supplementary or auxiliary information must be correlated with the variable under study. This method of estimation is usually appreciated and generally gives more precise estimates regarding population characteristics as compared to simple estimation method. The supplementary information used in estimation is usually referred to as **Benchmark Variable or Auxiliary Variable.**

When information on auxiliary variable is available certain methods of estimation are available. Two most important methods of estimation are ratio and regression method.

The term ratio estimation applies to a widely used class of estimation plan which incorporates the prior information including individual values, which is closely related to estimand for all units of the population. Such situations arise very frequently in practice, especially where the variables under consideration are the current level of some economic variable (such as retail sales) and the results of a survey giving some previous level of that variable.

Given a population *N* units (*Y*1, *X*1), (*Y*2, *X*2), …., (*Y*N, *X*N) the **population ratio** is defined as:

=, (6.1.1)

where  is variable under study and *Xi* is benchmark variable. If (*y*1, *x*1), (*y*2, *x*2), …., (*y*n, *x*n), are the values based on a simple random sample of size *n*, drawn from a population of *N* units, the **sample ratio** is defined as:

, (6.1.2)

The sample ratio given in (6.1.2) is generally used as an estimator of population ratio given in (6.1.1).

The **ratio estimator** or **classical (conventional) ratio estimator**, , of the population total *Y* is:

 (6.1.3)

The ratio estimator of the population mean can be written directly using (6.1.3) and is given as:

 (6.1.4)

Ratio estimator given in (6.1.2) to (6.1.4) is generally biased and consistent *r*. Ratio estimator has the following properties:

* It is generally biased, but the bias decreases as the sample size increases as in case of unbiased estimation and the distribution of r tends to normality (this is a property of consistent estimator), and
* It has smaller mean square error than that of unbiased estimator, if there is high correlation between benchmark variable and estimand.

Ratio estimators can be formed for any selection plan, but in this chapter attention will be confined to simple and stratified random sampling.

We will also discuss, in this chapter, the Ikeda-Midzuno-Sen procedure which makes the ratio estimator an unbiased estimator.

More over when the relation between estimand and benchmark variable is linear and does not pass through the origin, we use another method of estimation, called **Regression Estimator**. Ratio and regression estimators are identical when the regression line passes through the origin. In fact ratio estimator is a special case of regression estimator.

The regression estimator may be defined as follows:

If (*y*1, *x*1) (*y*2, *x*2), ..., (*yn*, *xn*) are the sample values, based on simple random samples, drawn from a population of *N* units, then the classical regression estimator  of population mean  is defined as:

 (6.1.5)

where *byx* is the regression coefficient of *y* on *x* and is given as:

 (6.1.6)

Ratio and regression estimator will also be discussed when two auxiliary variables are used.

**6.2 expectation and mean square error of   
ratio and ratio estimator**

In this section expectation and variance (mean square error) of ratio and ratio estimators are derived.

**6.2.1 Ratio**

**THEOREM (6.1)**

For large *n*, expected value of ratio estimator is approximately equal to *R* i.e.

, (6.2.1)

with approximate variance(m.s.e)

 (6.2.2)

### proof

We know that 

Now

 (6.2.3)

Since n is large, in the denominator may be taken as then (6.2.3) becomes

 (6.2.4)

Taking expectation of both sides of (6.2.4), we get

,

as . Hence E(r) ≈ R **◊**

In order to derive mean square error, the expectation of square of both sides of (6.2.4) is taken

.

Using the concept of simple random sampling [Theorem (2.2)], we can write

◊

**Remarks (i):**

An approximate unbiased variance estimator of (6.2.2) may be written directly using the sample analogy as:

 (6.2.5)

If  is not known then (6.2.5) can alternatively be (replacing  in the denominator) written as:

. (6.2.6)

For calculation purpose, the most convenient form of (6.2.5) is

. (6.2.7)

Rao and Rao (1971) compared (6.2.5) and (6.2.6) and found that (6.2.6) is frequently less biased.

**Remark (ii)**

The expression (6.2.2) may also be written alternatively as:

, as 

, (6.2.8)

, (6.2.9)

where  is the correlation coefficient between *X* and *Y* i.e. = 

**Remark (iii)**

The equation (6.2.8) may also be put as

,

. (6.2.10)

where *CY*, *CX* are coefficient of variation for *Y* and *X* respectively and *Cyx* are coefficient of co-variation.

**6.2.2 Ratio Estimator**

The difference between ratio and ratio estimator is only a multiple of *X*, which is a constant multiplier. It is clear that in the analysis of sampling error  is related to that of ratio, r. It can be easily proved analogues to (6.2.1) that

 (6.2.11)

The MSE of  can be easily obtained by multiplying (6.2.2), (6.2.8), (6.2.9) and (6.2.10) with .

The approximate MSE of is given as:

 (6.2.12)

 (6.2.13)

 (6.2.14)

 (6.2.15)

 may be written in a straight forwarded manner and is given as:

 (6.2.16)

or

, (6.2.17)

where =.

An approximate variance (m.s.e) estimator of (6.2.12) is

 (6.2.18)

Note that approximate variance (m.s.e) estimator for (6.2.13) and (6.2.14) may be obtained by replacing and r for R, in the respective expressions. An approximate variance estimator of (6.2.15) may be written directly and is:

 (6.2.19)

For practical purposes the most convenient form for (6.2.19) is

 (6.2.20)

(6.2.3) may also be written as [(Hansen, Hurwitz and Madow (1953) Vol. II   
Chapter 5)] to the correct to order *n*-1

 (6.2.21)

An approximate variance estimator of (6.2.21) is

 (6.2.22)

Expression (6.2.17) may also be derived through the following concept. To discuss this estimator designed to suit different situation we introduce the following notations. We write

 (6.2.23)

Consider the classical ratio estimator

 (6.1.4)

Using the notations given in (6.2.23) and neglecting second and higher order terms we get

 

Now



The mean square error of  is:







Thus the mean square error (MSE) of  is given by

 (6.2.17)

**6.3 SOME RECENT DEVELOPMENTS ON RATIO ESTIMATORS**

The ratio estimator given in the previous sections was based on the information of a single auxiliary variable. It was also shown in section 6.5, what condition the ratio estimator is more efficient that unbiased estimator of simple random sampling. There we had shown that the efficiency of ratio estimator is dependent upon correlation between estimand and auxiliary variable. The efficiency of ratio estimator can be substantially increased by using more auxiliary variables.

Certain modifications of the ratio estimator with more than one auxiliary variable have been given in literature from time to time. Some of these modifications are given in the following subsections.

**6.3.1 Modification of Classical Ratio Estimator-I**

Chand (1975) developed a chain ratio type estimator in the context of two phase sampling. It seems sensible to study the possibility of adapting it to the new situation although the force of its argument is somewhat lost in the single phase case. Samiuddin and Hanif (2006) has proposed following ratio cum ratio estimator using two auxiliary variables is:

 (6.3.1)

The mean square error of (6.3.1) is derived in the following theorem.

**THEOREM (6.2)**

The mean square error of (6.3.1) is



(6.3.2)

**PROOF**

Using the concept given in (6.2.23) we get

 (6.3.3)

Ignoring second and higher order terms we get

 (6.3.4)

The mean square error of  is:

= (6.3.5)

Using (6.3.4) in (6.3.5) we get



Applying expectation we get



On simplification we get



(6.3.2)

****

 (6.3.2)

where is classical ratio estimator.

**6.3.2 Modification of Ratio Estimator-II**

Samiuddin and Hanif (2006) have proposed another ratio type estimator with two auxiliary variables. The estimator is

 (6.3.6)

The mean square error of  is given in the following theorem.

**THEOREM (6.3)**

The mean square error of  given in (6.3.1) is:



 (6.3.7)

**PROOF**

Using the concept given in (6.2.23)



Expanding and ignoring second and higher order terms we get

 (6.3.8)

Similarly

 (6.3.9)

Using (6.3.8) and (6.3.9) in (6.3.6) we get



On simplification



The mean square error of  is



or

 (6.3.10)

Using (6.3.8) and (6.3.9) in (6.3.10) we have:



 (6.3.11)

To get the optimum value of “a” we partially differentiate (6.3.11) w.r.t. “a” and equating it to zero we get



The optimum value of “a” is





or

 (6.3.12)

Taking the square of (6.3.11)







Applying expectation the mean square error will be



 (6.3.13)

Putting the value of “a” from (6.3.12) in (6.3.13) and on simplification we



 (6.3.7)

Since  are special cases of  therefore we conclude that  and .

**6.4 BIAS OF THE RATIO ESTIMATOR**

Since *E*(*r*) is approximately equal to *R*, bias in the ratio estimator exist and may be obtained as under:

==.

Expanding the last term of above expression by Taylor’s expression and neglecting the higher order terms we get:



Taking the expectation we get:

Bias ≈

=

=

= (6.4.1)

=

=

= (6.4.2)

Bias is negative, zero or positive provided .

(6.4.1) may alternatively be put as

 (6.4.3)

Bias = 0, if *R* =  which is a condition that regression line *y* on *x* is a straight line through the origin.

Hartley and Ross (1954) expressed the bias in the following form:

 (6.4.4)

This can be proved as;



or 



This does not depend on the size of sample.

The bias in  may be estimated from (6.4.2) and (6.4.3) by multiplying it with *X* this yield:

Bias (6.4.5)

and Bias (6.4.6)

**6.5 UNBIASED RATIO ESTIMATOR**

In the previous sections we have seen that, under simple random sampling, classical ratio estimator is biased. Lahiri (1951) has shown that the classical ratio estimator can be made unbiased by changing the selection procedure. Midzuno (1950) and Sen (1951) proved the same result. Lahiri suggested that the first unit was selected with probability proportional to the aggregate of the size (PPAS) or with probability proportional to, and the remaining *n –* 1units with equal probability and without replacement. Midzuno (1951) simplified this procedure as “the first unit is selected with probability proportional to Xi (measure of size), and the remaining (*n* – 1) units with equal probabilities and without replacement”. This idea was introduced by Ikeda (1950) – reported by Midzuno (1951). This sampling scheme has striking resemblance to the simple random sampling without replacement. In fact, it may be viewed as a generalization of the simple random sampling when extra information on the population is available.

The idea is illustrated below:

Suppose we have a population of N units. The probability that ith unit is first selected with probability proportional to size and remaining (n – 1) units are selected with equal probabilities.

.

Similarly probability that *j*th unit is first selected with probability proportional to size is and subsequent (*n* – 1) units with equal probability and without replacement is given as:

,

The probability of sample, *P*(*s*), for the two selections is therefore:

.

Since there are n such selection therefore the probability of the selection of the sample will be

 . (6.5.1)

. (6.5.2)

The unbiasedness of ratio estimator under this selection procedure is proved int the following theorem.

**THEOREM (6.4):**

Classical ratio estimator is unbiased under Ikeda-Midzuno–Sen–Lahiri selection procedure with variance

 (6.5.3)

where is sum over all possible samples.

**PROOF**

Taking the expectation of (6.1.3) we have

,

where *P*(*s*) is the probability of the sample..Putting the value of *P*(*s*) from (6.5.1) we have



On simplification we get

◊ (6.5.4)

The variance of  may be derived as;



*Var*()

Substituting the value of *P*(*s*) from (6.5.1)

.◊ (6.5.3)

Note that the. This is very strong property and will be referred to as **Ratio Estimator Property***.*

Analogously to the ratio estimator of population total the unbiasedness and variance of ratio estimator of mean is proved in the following theorem:

**THEOREM (6.5)**

The ratio estimator of mean is unbiased with variance

 (6.5.6)

**PROOF**

Taking the expectation of (6.1.4) we get



Using (6.5.2) we have

◊ (6.5.6)

Proceeding in the same way we can derive the variance expression of, i.e.

 (6.5.7)

The unbiased variance estimator is given in the following theorem:

**THEOREM (6.6)**

An unbiased estimator of  is

 (6.5.8)

 (6.5.9)

**PROOF**

It may be proved that. For this consider:



=  (6.5.10)

and



 (6.5.11)

Hence



Similarly we can show that an unbiased estimator of population total is:

 (6.5.12)

**6.6 RATIO ESTIMATOR UNDER SUPER POPULATION MODEL:**

Consider all estimators  of *Y* that are linear functions of sample values *yi*, that are of the form

 (6.6.1)

where  does not depend on  though they may a function. The choice of the is restricted to those that give unbiased estimation of. The estimator with the smallest variance is called **best linear unbiased estimator(BLUE)**.

Consider linear stochastic model as:

 (6.6.2)

where ε*i* are independent of the and  are > 0. The  (*i* = 1,2,…,*N*) are known. The model is the same that was employed by Cochran (1953), which appears to have been originated by H.F. Smith (1938). Useful references to this model are Cochran (1953, 63, 77), Brewer (1963b), Godambe and Joshi (1965), Hanif (1969) Foreman and Brewer (1971), Royall (1970), Brewer and Hanif (1983) Cassel, et al (1976), Isaki and Fuller (1982), Hansen, Madow and Tepping (1983), Samiuddin et al (1992)and many others.

Brewer (1963b) defined an unbiased ratio estimator under the model (6.8.2). He used the concept of unbiasedness which was different from that given in randomization (design – based) theory. Brewer and Royall regarded an estimator  (estimated population total) is unbiased if  in repeated selections of the finite population and sampled under the model. Under model (6.6.2) Brewer (1963b) proved that the classical ratio estimate was model – unbiased and is best linear unbiased estimator for any sample [random or not] selected solely according to the values of the. This result hold goods if the following line conditions are satisfied;

1. The relation between estimated () and benchmark () is linear and passes though the origin.
2. The  about this line is proportional to xi.

In the following theorem we have shown that the ratio estimator is BLUE under the Linear Stochastic model (6.6.2).

**THEOREM (6.7):**

Under the model (6.6.2) classical ratio estimator is unbiased with variance

= (6.6.3)

**PROOF:**

We know that

 (6.6.4)

Using model (6.6.2) we have



Since = 0 then we have:

 (6.6.5)

We also know that

 or  (6.6.6)

Now



Now if  (6.6.7)

Therefore we say that  is model unbiased if

 (6.6.8)

The variance of , is:

 (6.6.9)

Now 

Using the condition of model we have:

 (6.6.10)

Using (6.6.2), (6.6.5) and (6.6.10) in (6.6.9), we will have

 (6.6.11)

Let us for simplicity we assume  then (6.6.11) will be:

 (6.6.12)

We can minimize  w.r.t. *ci*. For this the Lagrange’s multiplier is given as:



Differentiating unconditionally with respect to *ci*, we get.



or (constant)

We know from (6.6.7) that 

or



Hence 

So the BLUE is ratio estimator



Since  and 

Divide  into sample and non-sample values we have

or 

Squaring and taking the expectation





Substituting the value of Var(xi), we have:



 (6.6.3)

Using all these assumptions a model-unbiased estimator  from the sample may be easily proved as

. (6.6.13)

Putting this value of  in (6.6.4) a model-unbiased variance estimator is

 (6.6.14)

This model based unbiased estimator is not only superior to  but is the best of a whole class of estimators. For details see Brewer (1963b, 1979), Royall (1970), Royall and Herson (1973) and Samiuddin et al (1978).

**6.7 COMPARISON of  AND **

The performance of ratio estimator over the mean per unit estimator can be easily decided by comparing their variance. The condition under which ratio estimator is more precise than the mean per unit estimator is derived below:

We know that

 (6.2.8)

and  (2.4.1)

The ratio estimator will be more precise if  or if



or 

Hence the efficiency of ratio estimator with respect to simple (unbiased) estimate is not only depending on the high correlation but also depends on the coefficient of variation of variables. The benchmark variable must have high coefficient of variation than estimand. Further comparison under a stochastic model is also given.

It is an established fact that the choice of a suitable sample plan is central to the design of a sample survey. Sample design can be regarded as comprising separate selection and estimation procedures, but the choices of these are so interdependent that they must be considered together for virtually all purposes. Some times the nature of the sample plan is determined by circumstances, but usually the designer is faced with a choice, and frequently it is obvious which of a number of possible plan will be most efficient in terms of minimum sample error for given cost( or vice versa). Standard sampling theory using imputed values for such quantities as the means, variances, and correlation coefficient of the (finite) population, or strata or clusters within it, can often indicate which design is most efficient. Sometimes, however, this is not so. A well-known example is the comparison between classical ratio estimation using unequal probabilities. To obtain a straight forwarded answer in this case, Cochran (1953) made use of a certain super population model (6.6.2) which is intuitively attractive and appears to have some empirical basis. The purpose here is to compare classical ratio estimator and unbiased estimation method using equal probabilities and using large scale sample results which can be obtained using generalization of model. Comparison for probability proportional to size will be discussed in Chapter 7, 8 and 9. The stochastic model used here for the purpose of comparing efficiencies is given in (6.6.2).

In the following theorem we have given the condition under which ratio estimator is more precise then the unbiased estimator of population total is simple random sampling.

**THEOREM (6.8):**

Under linear stochastic model (6.6.2) ratio estimator will be more efficient than unbiased estimator

**PROOF**

We know that:



Putting the value of  from (6.6.2) we have



= (6.7.1)

Also (6.7.2)

or  (6.7.3)







 (6.7.4)

as cross product term is equal to 

Now the first term of (6.7.4) using (6.6.2) is





 (6.7.5)

Similarly



or

 (6.7.6)

Using (6.7.5) and (6.7.6) in (6.7.4) we get:

 (6.7.7)

Now consider the ratio estimator as:

 (6.1.3)



 (6.7.8)

Now







 (6.7.9)

Now





 (6.7.10)

Now we know that



Therefore  (6.7.11)

Comparing (6.9.7) and (6.9.1) we have:



So Ratio Estimator will always be more efficient

Foreman and Brewer (1971) used the following model 

With the same assumption given in (6.7.2) they compared various method of estimation and proved that ratio method of estimation is more efficient than unbiased estimation method provided | α | < | β*X* |.

**6.8 HARTLEY-ROSS UNBIASED RATIO ESTIMATOR**

Hartley and Ross (1954) has proposed a modified ratio estimator in simple random sampling. The modified estimator proposed by them is:



The expectation and variance of is obtained in the following theorem:

**THEOREM (6.9):**

In simple random sampling without replacement an unbiased estimator of R is

, (6.8.1)

where [average of the ratio] and for large sample the variance of  is:

. (6.8.2)

where .

#### PROOF

We know that (using the concept of simple random sampling)







Therefore

 ◊ (6.8.3)

Also from simple random sampling we have:



or



Cov (ri xi) can be estimated by this concept, i.e.





so

 (6.8.4)

Taking expectation of (6.8.1) and using (6.8.3) and (6.8.4)





Hence 

**Note:** Since  which is not easy to calculate so in large sample survey is unlikely to be used, although it is an unbiased estimator.

The corresponding unbiased ratio estimator  for population total Y is

 (6.8.5)

 (6.8.6)

Now for large n (6.8.1) takes the form





By the law of large number the random variable  converge in probability to. Hence the limiting distribution of



is the same as the distribution of



Thus



 (6.8.7)

Note that.

From (6.2.9) (ignoring correction factor) and (6.8.7) we have:





where is regression coefficient of *y* on *X*.





We can conclude that *Var*(*r*) is less than *Var*(*rHR*) if β is closer to *R* then to. This result was given by Goodman and Hartley (1958). Note that the efficiency of rHR over classical ratio estimator depends upon the nature of the relation between *Y*, *X* and *R*. Also note that it does not provide satisfactory variance estimator, as approximation is applied as in case of classical ratio estimator. Goodman and Hartley (1958) also derived exact formula for variance.

**2.2.2 Prasad’s (1989) Ratio Estimator**

Prasad (1989) proposed a ratio -type estimator for estimating of population meanON THE LINES SUGGESTED BE Searl(1964). The proposed estimator is

, (2.2.5)

where  is constant

In this ratio type estimator an auxiliary variable  correlated with *y* is measured for each unit in the sample. The population total *X,* like classical ratio estimator assumed to be known. In practice,  may be value of ** at some previous occasion when a complete census or a sample survey was taken or it may concurrently be measured along with, *y* on each unit in the sample (i.e. weight ** of an orange in the sample may be measured while measuring the content of vitamin C in the orange). The aim, like classical ratio estimator, is to obtain increased precision by taking advantage of the correlation between *y* and *x*. We know that ratio method of estimation of  is based on the sample mean  and of ** and *X*. Thus the efficiency of the ratio method of estimation is very much dependent on the efficiency of andas an estimator of and respectively..

We know that in many sample surveys reduction in mean Square error, even by a very small amount, plays an important role to increases efficiency significantly of the overall estimators. For example, in stratified random sampling with ** strata the mean square error for the estimator of the individual strata mean (Cochran 1977, p-165). Thus, at **, the number of strata, is large, the over mean Square error could be very large, even a slight reduction in the mean square every for the individual stratum mean estimator, would increase the efficiency of the overall population mean estimator significantly. We define sample ratio r, as:

Let  . (2.2.6)

Subtracting R form both sides and simplifying we get



  (2.2.7)

Expanding (2.2.7) we get





 (2.2.8)

We know that  (2.2.9)

As the optimum value of k = (1.6.5)

and

  (2.2.10)

Using (2.2.9) and (2.2.10) in (2.2.8) we get



 





 (2.2.11)

 for all therefore

> 

Putting the value of k from (1.6.5) in (2.2.11) we get



if  , If then 

**6.9 REGRESSION ESTIMATOR**

Theoretically, it has been established, in general, that the linear regression estimator is more efficient than the ratio and product estimators except in the case where the regression line passes through the neighborhood of the origin, in that case the efficiencies of these two estimators are almost equal. However, due to strong intuitive feeling statisticians are more interested towards the use of classical ratio and product estimators when the regression line does not pass through the origin. Srivastava(1967), Walsh (1970), Reddy (1973,74), Gupta (1978), Sahai (1979), Vos (1980), Adhvaryan and Gupta (1983) etc. have proposed certain estimators which under certain realistic conditions are more efficient than mean per unit. The classical ratio estimator, product estimator is as efficient as regression estimator.

The ratio estimator and its modifications given in the previous sections can be efficiently used for estimation of population characteristics in presence of auxiliary information.

One other method, that also utilities the auxiliary information, of estimation is the regression method of estimation. This method of estimation, in survey sampling, has proved to be more efficient than the ratio method of estimation. Specifically, the estimator of population mean in this method, called the regression estimator is defined as:

 (6.9.1)

The expectation and variance of (6.9.1) is given in the following theorem.

**THEOREM (6.10):**

Under simple random sampling the regression estimator is unbiased with variance:

 (6.9.2)

###### PROOF

We know that

 (6.1.5)

or



Taking the expectation, we have



Hence ◊. If .

Note that the unbiasedness does not depend on the value of byx.The variance of  is



Substituting the value of  from (6.9.1) in the above expression and   
re-arranging the terms



Using the idea of simple random sampling we have:





 (6.9.3)

Since  is an unbiased estimator for all values of *byx*, the question arises, what should be the best value of *byx* so that  minimum. For this we find the partial derivative of (6.9.3) w.r.t. *byx* and equate it to zero as:

 (6.9.4)

or

 (6.9.5)

where β*YX* is population regression coefficient of y on *X*.

Using (6.9.5) in (6.9.2) and on simplification, we get

 (6.9.6)

which takes minimum value.

The minimum variance of  may also be obtained by using the concept given in (6.2.23) as under:

Consider the classical regression estimator

, (6.1.5)

where *byx* is the sample estimate of, the population regression coefficient of *y* on *x*. Using (6.2.23) in (6.1.5) we get



or





Applying expectation we get



Putting the value of  we get



On simplification we get



which is the same as (6.9.6)

Also it is known that.

The estimated population total can be written as:

 (6.9.7)

The variance of  is

 (6.9.8)

Also (6.9.9)

If we compare (6.2.14) and (6.9.8), we get:



and say



iff  (6.9.10)

which is always true.

Both estimators become identical if  which is the optimum value of *byx* as it minimizes the . Comparing simple random sampling with regression estimate we have:

 (6.9.11)

Hence  is always more efficient than. Both are identical if  = 0.

**6.10 RECENT DEVELOPMENTS IS REGRESSION ESTIMATOR**

Certain developments have been made in regression estimator from time to time. Some of these developments are given in the following:

**6.10.1 Mohanty’s Estimator for One-Phase Sampling   
[Using two Auxiliary Variables]**

Mohanty has proposed a regression cum ratio estimator. The estimator utilized information of two auxiliary variables and is given as:

 (6.10.1)

The mean square error of (6.10.1) is given in the following theorem.

**THEOREM (6.11):**

The mean square error of  (6.10.1) is given as:

 (6.10.2)

**PROOF**

Using (6.2.23) in (6.10.1) we get





Expanding and ignoring higher terms we have:

 (6.10.3)

Squaring (6.10.3) and applying expectation the mean square error of  is:





Applying expectation and substituting the value of 



On simplification we get



(6.10.4)

or

 (6.10.2)

**6.10.2 Revised Regression Estimator**

Samiuddin and Hanif (2006) has given a revised regression estimator as:

 (6.10.5)

The estimator given in (6.10.5) is a simple extension (6.9.1) with one additional auxiliary variable.

The mean square error is derived below:

**THEOREM (6.12):**

The Mean square error of (6.10.5) is:

 (6.10.6)

**PROOF**

If we consider the regression estimator of *Y* on *X* and *Z* and use sample information we may proceed as follows:

Let

(i= 1,2,……n) (6.10.8)

The Least Square (L.S.) estimate of  are

,  and. (6.10.9)

Using these we can write the estimated  as

 (6.10.10)

Consequently an estimator of  is

, (6.10.11)

where  and  are to be determined so that  is minimum. Using (6.2.23) we get

 (6.10.12)

 (6.10.13)

Also  and  are determined so that  is minimum. This leads to

or 

or 

Solving the above we have:

 (6.10.14)

and

 (6.10.15)

Now



Using  we get









 (6.10.16)

Substituting the value of  and  in above equation and on simplification we get



 (6.10.7)

where  (6.10.17)

**6.10.3 Modification of Mohanty’s Estimator-1**

Samiuddin and Hanif (2006) has given certain modifications of Mohanty’s estimator. The first modification given by Samiuddin and Hanif (2006) has the form

 (6.10.18)

where  and 

The mean square error is given in the following theorem.

**THEOREM (6.13):**

The mean square error of modified Mohanty’s estimator is:



(6.10.19)

where

 (6.10.20)

 (6.10.21)

and

 (6.10.22)

**PROOF**

Let



 (6.10.23)

or

 (6.13.24)

In order to find the optimum value of *a*, square (6.10.23) and differentiate with respect to a and equate it to zero to get:

 (6.10.25)

and



(6.10.26)

Now using (6.10.25) and will be



Using above two equations in (6.10.23) and squaring and taking expectation the mean square error is







Now



Applying expectation we get



Interchanging  and 



Now







Applying expectation and putting the value of and we get







or





**6.10.4 Modification of Mohanty’s Estimator-2**

Another modification of Mohanty’s estimator is straight forward. Samiuddin and Hanif (2006) has shown that the estimator

 (6.10.27)

can also be used for estimation of population total. The mean square error of (6.10.27) is derived in the theorem below:

**THEOREM (6.14):**

The mean square error of (6.10.27) is

(6.10.28)

**PROOF**

We have to determine such that  is minimized.

Using (6.2.23) in (6.10.27) we will have



Expanding and ignoring higher terms we have:



The mean square is

 (6.10.29)

The constant  is to be determined so that is minimum.

The optimum value of can be determined by differentiating (6.10.29) and equating to zero. This gives



or

 (6.10.30)

Squaring the right hand side of (6.10.29) we get



Applying expectation





Substituting the value of  in the above equation we get



On simplification we get



or





The value of  which minimizes is given as:



The estimated  is given as

.

If we substitute the value of  in (6.10.27) the estimator becomes:

 (6.10. 31)

If we compare (6.10.28) with, we notice that there is difference in only one term that is  in  is replaced by  in . Thus   and the reduction may be substantial if *Z* is closely related to *Y* so that  is small.

**6.10.5 Modification of Mohanty’s Estimator-3**

The third modification of Mohanty’s estimator is given by Samiuddin and Hanif (2006) is:

 (6.10.32)

The estimator is a slight modified version of. The mean square error of  is given in the theorem below:

**THEOREM (6.15):**

The mean square error of  is

 (6.10.33)

**PROOF**

Since the different between  and  is only the interchange of *X* and *Z* therefore  will be written on the lines with (6.10.28) i.e.

 (6.10.34)

Substituting the value of  from (6.10.34) we get an estimator 

 (6.10.35)

Thus the MSE may be written on the lines of as:



**6.11 Unbiased Variance Estimator**

An unbiased variance estimator of (6.12.2) may be written in a straight forward way as

 (6.11.1)

For computation purposes the most convenient form is:

 (6.11.2)

An unbiased estimator of (6.11.7) [using the concept of simple random sampling] is

 (6.11.3)

Unbiased variance estimators of (6.11.8) and (6.11.9) are

 (6.11.4)

and

 (6.11.5)

respectively.

# EXAMPLE (6.1)

##### (6. 1)

The population of Greece for 69 Urban areas is known from 1941 population census is *X* = 13,559 in hundrads. From this 20 urban areas are selected at random using simple random sampling without replacement and the population for 1941 for these areas where noted from the census record and the present population for these 20 areas were obtained through field work to estimate the present population of 69 areas. Estimate the present population using ratio estimation method and simple unbiased estimating method. Compare the efficiency of these two methods (Raj 1972): Data are given in Table 6.1.

**SOLUTION:**

We have following information:





**Table 6.1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***x*** | ***y*** | ***x*2** | ***y*2** | ***xy*** |
| 122 | 126 | 14884 | 15876 | 15372 |
| 290 | 257 | 84100 | 66049 | 74530 |
| 70 | 108 | 4900 | 11664 | 7560 |
| 141 | 185 | 19881 | 34225 | 26085 |
| 497 | 421 | 247009 | 177241 | 209237 |
| 130 | 148 | 16900 | 21904 | 19240 |
| 87 | 119 | 7569 | 14161 | 10353 |
| 198 | 338 | 39204 | 114244 | 66924 |
| 347 | 368 | 120409 | 135424 | 127696 |
| 151 | 221 | 22801 | 48841 | 33371 |
| 83 | 121 | 6889 | 14641 | 10043 |
| 153 | 151 | 23409 | 22801 | 23103 |
| 156 | 224 | 24336 | 50176 | 34944 |
| 327 | 410 | 106929 | 168100 | 134070 |
| 304 | 295 | 92416 | 87025 | 89680 |
| 139 | 177 | 19321 | 31329 | 24603 |
| 623 | 790 | 388129 | 624100 | 492170 |
| 150 | 176 | 22500 | 30976 | 26400 |
| 217 | 236 | 47089 | 55696 | 51212 |
| 91 | 129 | 8281 | 16641 | 11739 |
| 4276 | 5000 | 1316956 | 1741114 | 1488332 |

*x* = population for 1941; *y* = present population

**(i) Ratio estimation method**

= *r*.*X* = (1.16932) (13559) = 15885

 = 543874.826

*S*.*E*. = 737.4786

**(ii) Simple estimation method**

= 436922.193

= 2090.364

The efficiency of classical ratio estimator  over simple estimation method  is



From the data given in example 6.1 estimate total number of persons using the method of regression estimate and find the standard error of your estimate. Compare your estimated values with the one that obtained from ratio estimate.

# SOLUTION

We know that



Total number of persons

 



We can easily compare, andas

|  |  |  |  |
| --- | --- | --- | --- |
|  | Regression | Ratio estimate | Simple estimate (SRS) |
| Estimated total | 16007 | 15885 | 17250 |
| Estimated variance | 485031.4398 | 543874.826 | 4396622.193 |
| Standard error | 696.442 | 737.4786 | 2090.364 |

We can see that:

<  < 

**6.12 DIFFERENCE ESTIMATOR**

We have already seen in Section (6.8) that ratio estimator is best linear unbiased estimator if the relation between estimand (*Yi*) and benchmark (*Xi*) variables are linear and passes through the origin, i.e. *Yi*– *kXi* = 0, where *k* is constant. In practical life such type of relation is not always possible. There might be a situation when the relation between *Yi* and *Xi* of the form *Yi – k Xi = a*, where a is constant. We wish to estimate  and . When *xi* and *yi* are correlated then estimator  may be improved by introducing a factor known as difference **factor**. If it is assumed that there is a unit change in *y* when a unit charge is made in *x*, then simple difference function may be introduced as;

 (6.12.1)

It is further assumed that x and y have equal variances. A more generalized form of (6.12.1) may be defined as:

, (6.12.2)

where  and k are known.

**THEOREM (6.16):**

The difference estimator (6.12.1) is an unbiased estimator of population mean with variance is:

 (6.12.3)

**PROOF**

Taking the expectation of (6.12.1)



Using the concept given in the simple regression on can write



Taking the square and applying expectation on both sides



Using the idea of simple random sampling we have:

 (6.12.12)

An optimal value of *k* can be obtained by minimizing (6.10.3). For this we find partial derivative of (6.12.12) w.r.t. *k* and equate it to zero to get:

= 0 (6.12.4)

Therefore, 

Putting the value of k in (6.12.4), we have



If k = R, then



which is variance of  for ratio estimator.

If K = 0, then 

An unbiased estimator of (6.17.3) is

 (6.12.5)

The difference estimator is superior to, if 

or 

Therefore *k* lies between 0 and 2 β. If *k* is outside the range then  is superior to.

**6.13 REGRESSION ESTIMATE AS MODEL-UNBIASED**

The design based paradgim for regression estimator, discussed in the previous sections, provide ground for estimation of population characteristics when information about a population is available. It often happens that a population under study is a realization from some super population model.

In section 6.8 we have given rigorous treatment to ratio estimator under model (6.8.2). We have shown there that under model (6.8.2) the ratio estimator is BLUE. In the following we have given rigorous treatment to regression estimator under the model

 (6.13.1)

where .

Parallel to ratio estimator, the regression estimator has very important property under super-population model and is proved in the following theorem.

# THEOREM (6.17):

Under model (6.13.1), is unbiased for any sample size with variance:

 (6.13.2)

**PROOF**

We know that

 (6.13.3)

Substituting the value of *yi* from model (6.13.1) and on simplification we get

 (6.13.4)

Since  is unbiased estimator of  which is zero. Thus  is distributed normally about zero mean in repeated samples. This is of the order  as standard error of sample covariance is of the order. Also is of the order unity. Therefore, (b – β) is of the order.

Under the model (6.13.1) we have:



and

 (6.13.5)

We also know that



or



Substituting the value of  from the model and (*b – B*) from (6.13.4) and on simplification

 (6.13.6)

Taking expectation



Since

, therefore 

Hence  is model unbiased.◊

Also squaring (6.14.6) and taking expectation under the model.





or

 (6.13.2)

Actually under super population model regression estimator is Best linear unbiased predictor (BLUP)

**6.13.1 Unbiased Variance Estimator**

The unbiased variance estimator under super population model is derived in the following:

For unbiased variance estimator we proceed as under:

Let 

then 

From this we have







We have proved that b - β is of the order. Hence we left with



or





Since , Hence



is model unbiased.

**6.13.2 Bias of Regression estimator**

If we use a simple random sample then linear regression estimate is biased i.e., the bias of the regression estimate is trivial and it decreases as size of the sample increases. The bias comes in  due to two reasons.

1. β is estimated by the ratio of  to that of 
2. It involves the product of b and which are estimates.

The amount of bias in regression estimator is obtained in the following:

**THEOREM (6.18):**

The bias in regression estimator using simple random selection procedure is.

**PROOF**

Suppose



and



where



Putting these values in (6.11.1) we have



Further



Since E(ε) = E(ε1) = 0. Therefore,



 (6.13.8)

If  then regression estimate reduces to ratio estimate then bias in regression estimates takes the form

 (6.13.3)

**6.14 RATIO AND REGRESSION ESTIMATORS   
IN STRATIFIED SAMPLING**

So far we have discussed ratio and regression estimator under simple random sampling design. Efficiency of these estimators increases tremendously if they are used in stratified sampling design. In the following we have given the two estimation methods in stratified random sampling.

**6.14.1 Ratio Estimator in Stratified Random Sampling**

Ratio estimation method may also be applied to stratified random sampling in two ways, namely, **stratum by stratum** (Separate Ratio Estimate) and **across stratum** (Combined Ratio Estimate).

The stratum by stratum ratio estimator  is

 (6.14.1)

This is an aggregate over strata to yield an estimate of population aggregate value, *Y*. Alternatively, a ratio estimator may be expressed directly in terms of   
*r* and the population benchmark *X*. This is called across-stratum ratio estimator and is expressed as:

 (6.14.2)

The MSE. of stratum by stratum ratio estimator may be written in a straightforward manner via the concept of stratified random sampling.

 (6.14.3)

 (6.14.4)

Approximate variance estimators of (6.10.3) and (6.10.4) are

 (6.14.5)

 (6.14.6)

The M.S.E. of  is



 (6.14.7)

 (6.14.8)

Approximate variance estimator of (6.13.7) and (6.13.8)



(6.14.9)

and

 (6.14.10)

respectively.

The difference of (6.13.4) and (6.13.8) given below, is worth noting

 

 (6.14.11)

Now the second term on the right hand side is usually smaller under the situation when the ratio estimator is applicable, it is obvious that the first term is always positive. This indicates that  may be smaller than that of. The bias of stratum by stratum ratio estimator is the sum of the biases associated with the estimate of each stratum, a decreasing function of *nh­­* while the bias of the across stratum ratio estimator is a similar decreasing function of n. The general conclusion that could be drawn from the above is that stratum by stratum ratio estimator is recommended only when the sample size is large within each stratum and when the sample size is small, across stratum method is recommended for applications.

A ratio estimate based on a stratified sample can be more efficient than one based on a simple random sampling of the same sample sizes except for stratum-by-stratum ratio estimates based on very small samples. Significance gains in the efficiency of ratio estimates can result, however, if sampling units are stratified on the basis of a suitable measure of size; other than that used as the ratio estimation bench mark variable.

Stratum by stratum ratio estimator is usually more precise than the corresponding across-stratum ratio estimator except when stratum sample sizes are small. A cross-stratum ratio estimators must be used when the population bench mark X is available but not the stratum bench marks *Xh*.

The advantage of estimating across strata is purely to reduce the bias introduced by ratio estimator. The bias is appreciable only when the sample sizes within a stratum are very small. The more the variation in *Rh*, the greater the gain for use of stratum by stratum estimators.

# EXAMPLE (6.2)

A pilot survey was conducted to estimate the total number of orange trees in Sargodha districts. The district was divided into two stratum according to the information provided by Revenue circle. A simple random sample was used to select the sample. Following information are available

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Stratum** | **Total number of village** | **Sample village** | **Total area under orange trees (Acres)** | **Area under orange trees (Acres)** | **Total number**  **of trees** |
| 1 | 632 | 8 | 25672 | 15.2, 3.2, 4.5, 10.8, 5.6, 9.8, 15.9, 21.6 | 940, 380, 415, 860, 410, 640, 815, 1120 |
| 2 | 730 | 10 | 20490 | 7.3, 1.2, 5.3, 2.1, 8.5, 7.8, 10.50, 2.8, 3.6, 5.1 | 480, 50, 230, 115, 500, 374, 516, 315, 315, 330 |

Estimate the total trees in 1362 villages using separate and combined ratio estimate and compare the efficiency of these two methods.

**SOLUTION;**

X1 = 25672, X2 = 20490, X = 46162 N1 = 632, N2 = 730, N = 1362

n1 = 8, n2 = 10, n = 18

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Stratum** | **Nh** | **Xh** | **nh** |  |  | **rh** |  |  |
| 1 | 632 | 25672 | 8 | 10.825 | 697.5 | 64.43 | 41.071 | 77728.6 |
| 2 | 730 | 20490 | 10 | 5.42 | 322.5 | 59.50 | 9.291 | 24716.1 |
|  | 1362 | 46162 | 18 |  |  |  |  |  |

= 496.51, xst = 7.93, = 6.4086, = 278.798, = 3.048

rc = 62.54, = 157.214 ρ1 = 0.9567, ρ2 = 0.888

Total number of trees

1. Separate Ratio Estimate () =  = 2873347
2. Combined Ratio Estimate=  = 2890276
3.  = 1744094849
4.  = 41762.36
5.  = 1116831797
6.  = 33419.033

The efficiency of  to 

= 

Here it is clear that the across ratio estimator is more efficient.

**6.14.2 Regression Estimator in Stratified Random Sampling**

We know that the regression estimator can be effectively used in Stratified random sampling design. To illustrate the idea we proceed as under:

 (6.14.12)

For hth stratum

 (6.14.13)

The regression estimate for estimated total  for separate (stratum by stratum) is

 (6.14.14)

with variance [may be written in straight forward way]

 (6.14.15)

 (6.14.16)

If bh = βh, then

 (6.14.17)

Similarly combined (across stratum) regression estimate is defined as

 (6.14.18)

*bc* is combined regression coefficient

The variance expression of  may also be written in a usual way i.e.

 (6.14.19)

If *bc = βc* then (6.14.19) gives minimum variance

 (6.14.20)

From (6.14.17) and (6.14.20) we have







 (6.14.21)

The R.H.S. of (6.14.21) is positive. Hence



This is true unless *βh* is the same for all the strata.

In fact no hard and fast rule can be given for the efficiency of separate regression and across regression estimator. However, only rough idea can be made. If regression is linear in all the strata and there is small variation in *byhXh* then across stratum regression estimate is advisable for application but if there is a large variation in *byhXh* from stratum to stratum then stratum by stratum regression estimate can be recommended for application. If the regression is non-linear, across stratum regression estimate is always better for practical application. Some further remarks regarding these are as:

1. Separate regression estimates is appropriate when true regression coefficient *βh* vary markedly from stratum to stratum.
2. Combined regression estimate is appropriate when βh is presumed to be the same in all the strata.
3. Separate regression estimate is more close to bias, when sample are small within the strata and variances have larger contribution from sampling error in the regression coefficient.
4. In combine regression estimate variance is inflated if the population regression coefficient differs from stratum to stratum.

##### EXAMPLE (6.3)

From the data given in Example 6.2 estimate the total number of trees by using separate and combined regression method of estimation and calculate the variance for each case.

##### SOLUTION

From the data given in Example 6.2, we have

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *b*1 | = | 41.6511 | *N*1 | = | 632 | *N* | = | 1362 |
| *b*2 | = | 47.6659 | *N*2 | = | 730 | *n* | = | 18 |
| *bc* | = | 49.2153 | *n*1 | = | 8 | *n*2 | = | 10 |
|  |  |  |  |  |  |  |  |  |
|  | = | 496.51 |  | = | 7.93 | *ρ*1 | = | 0.957 |
| *ρ*2 | = | 0.884 |  | = | 6.4087 |  | = | 2.995 |
| *sy*1 | = | 278.798 |  | = | 157.214 |  | = | 40.62 |
|  | = | 28.068 |  |  |  |  |  |  |

(a) Estimation of total trees









Comparison of may be seem in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RATIO | | REGRESSION | |
|  | Separate | Combined | Separate | Combined |
| Estimated Total | 2873347 | 2890276 | 2248616 | 2416579 |
| Estimated Variance | 1744094849 | 1116831797 | 798140486.2 | 736067487.7 |
| Standard Error | 41762.36 | 33419.033 | 28251.3803 | 27130.5637 |

**6.15 SOME OTHER RATIO ESTIMATORS**

Certain other ratio estimators has been proposed in literature. Some of these are given below:

**(a) Mickey’s Ratio Estimator:**

Mickey (1959) has given an unbiased ratio estimator. In this estimator the sample units are divided into k groups of size m each such that *n = mk*. An unbiased ratio estimator of population total *Y* is then:

, (6.15.1)

where

classical ratio estimator computed from the sample after omitting the jth groups for n = 2, reduced to.

**(b) Quenouille-Durbin Ratio Estimator:**

Quenouille (1956) has given a ratio estimator of the form:

 (6.15.2)

Cochran (1963) used modified from of  which is simplier and more convenient and is given as:



 (6.15.3)

**(c) Ratio Estimator Based on Group Sample Means:**

Another estimator is:

, (6.15.4)

where

.

reduces to  if n = 2

Tin (1965) used modified form of and is given as:

 (6.15.5)

In fact he used  in place of T in (6.11.4).

**(d) Pascual’s (1961) Ratio Estimator:**

Pascual (1961) has given another type of ratio estimator and is given as:

 (6.15.6)

**(e) Beale-Tin Ratio Estimator:**

These estimators are

 (6.15.7)

and

 (6.15.8)

Rao (1969) conducted empirical study to compare the efficiencies of these ratio estimators. He considered sample size *n* = 2, 4, 6, 8, 12 from the 20 natural and artificial populations. There is no definite conclusion which estimator is superior to another. It depends upon the nature of the population and also on sample size. In the example 6.5 and 6.6 given below we have only explained how the calculation for different estimators can be worked out.

**EXAMPLE 6.4**

A sample of 15 villages is drawn from a population of 50 villages and the following information for 15 villages is recorded *xi*, denotes the population for the previous year and *yi* denotes the population of the current year under estimate. Using this information estimate the total population of 50 villages. Compare this estimate with simple random sample. Also find the standard error in each case. [The true total, Σ*Xi* = 4746 and Σ*Yi* = 5431].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| xi |  | yi |  |  |  |  |
|  |  |  |  |  |  |  |
| 65 |  | 79 |  | 1.215 |  |  |
| 73 | 70.33 | 96 | 85.00 | 1.315 | 1.111 |  |
| 73 |  | 80 |  | 1.096 |  |  |
|  |  |  |  |  |  |  |
| 102 |  | 104 |  | 1.020 |  |  |
| 74 | 88.00 | 85 | 96.33 | 1.149 | 1.133 |  |
| 88 |  | 100 |  | 1.136 |  |  |
|  |  |  |  |  |  |  |
| 76 |  | 96 |  | 1.236 |  |  |
| 73 | 85.00 | 85 | 98.33 | 1.164 | 1.142 | 1.141 |
| 106 |  | 114 |  | 1.075 |  |  |
|  |  |  |  |  |  |  |
| 97 |  | 116 |  | 1.196 |  |  |
| 112 | 104.33 | 121 | 115.33 | 1.080 | 1.156 |  |
| 104 |  | 109 |  | 1.048 |  |  |
|  |  |  |  |  |  |  |
| 116 |  | 126 |  | 1.086 |  |  |
| 136 | 119.67 | 150 | 131.00 | 1.103 | 1.163 |  |
| 107 |  | 117 |  | 1.093 |  |  |

# SOLUTION

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | = | 1402 |  | = | 1578 |  | = | 17.002 |
|  | = | 137018 |  | = | 171378 |  | = | 1.133 |
|  | = | 93.467 |  | = | 105.2 | *r* | = | 1.126 |
|  | = | 152941 |  | = | 0.962 |  |  |  |
|  | = | 398.516 |  | = | 1.132 |  |  |  |
|  | = | 363.373 |  | = | 358.16 |  |  |  |

The estimated values of the populations using the estimators are given below:

|  |  |  |
| --- | --- | --- |
| **Estimators** |  | Estimated values |
| Classical Ratio |  | 5344 |
| Hartley – Ross |  | 5342 |
| Mickey’s |  | 5360 |
| Quenouille-Durbin |  | 5144 |
| Cochran |  | 5059 |
| Grouped Sample Mean |  | 5339 |
| Tin |  | 5336 |
| Beale-Tin |  | 5355 |

**6.16 JACK KNIFE AND BOOTSTRAP METHODS**

Tow very important methods of reducing bias are Jackknife and Bootstrap methods. These two methods are illustrated below:

**6.16.1 Jack Knife**

Jack Knife methods developed by Quenouille (1956) and Turkey (1958) for celebrating standard error, bias and confidence limits when there is no reliable information of the model available.

The idea is illustrated below:

Suppose a sample of observations ….. is available from a population with unknown parameter. Suppose the estimation of  is based on the sample. The probability distribution is unknown. If sample size is large, then the random variable  is approximately normal with mean zero and variance 1 where  and 

*E*(*T*) and *Var*(*T*) may be expanded as

*E*(*T*) = 

and

*Var*(*T*)= 

where  and , i=1,2….. are unknown.

One method is to divide the sample in to two equal sub-samples and to calculate  and  for  the bias and variance in  are

Bias =

Var(T) = 

Thus Bias = 

and  to the order.

In order to improve upon these bias and variance, Quenouille (1956) and Tukey (1958) and many others provide a technique called Jack knife.

Consider  to be a relevant estimator of  based on a sample of size *n*.

We define  to be another estimator of  based on n – 1 observations by excluding ith observation. Define a new estimator as:



where 

We find



This leads to 

where  then  is known as Jack Knife estimator of the Variance of.

The Bias is estimated as

Bias = 

The Jack Knife method provides confidence limit for  as



where is the point of the standard normal distribution, and



For example, consider  then

 reduces to



**Example 6.5**

Suppose a sample of size 5 is taken from a population with finite mean 4. The observations are 2, 3, 6, 5, 4.



Let all possible sample of size *n-1* when one is excluded be:

2,3,6,5 excluding 4 16/4=4

2,3,6,4 excluding 5 15/4=3.75

2,3,5,4 excluding 6 14/4=3.50

2,6,5,4 17/4=4.25

3,6,5,4 18/4=4.50



Bias = 0 this is so as we drew all possible samples of size 4.



**Example 6.6**

Suppose a sample of size 5 is drawn from a population with mean  and Variance . Suppose  is to be estimated. The observations are

2, 3, 6, 5, 4

**Solution**

Let the  be estimated by 

The calculations are made to find Jack Knife bias and Variance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** |  |  |  |  |
| 1 | 4 | 4.0 | 0 | 0 |
| 2 | 5 | 3.75 | 1.0 | 1 |
| 3 | 6 | 3.50 | 2.0 | 1 |
| 4 | 3 | 4.25 | -1.0 | 1 |
| 5 | 2 | 4.50 | -2.0 | 4 |
|  |  |  | = 0.0 | 10.0 |

where









5

**Example 6.7:**

Suppose *r* is the condition for pairs of 

**Table 1**

j    

1

2

3

4

5

*X* denotes scores on Test A

*Y* denotes scores on Test B

*r* denotes the sample correlation itself



Correlation coefficient when  removed from sample





Now we find r = 0.776 and =1.0352



**Example 6.8**

Suppose pairs of observations are collected, where

*Y* denotes average score on Test A

*Z* denotes average score on Test B

*j*     

1 52 3.4 0.89 5(0.78-0.89)=-1.62 1.43 -1.12

2 64 3.3 0.76 0.18 1.00 045

3 56 2.8 0.76 0.29 0.98 .71

4 58 3.3 0.78 0.00 1.04 0

5 67 3.4 0.78 0.63 0.93 1.46

6 58 3.1 0.78 -0.6 1.05 0.14

7 -0.11 1.06 -.29

:

:

15

*r* = 0.78, = 1.035

*r* Scale 

T Scale = = -.098



Bias = 

**6.16.2 Influence Function Explained in Robust Estimation**

….. from  *t* is an estimator of , we want to find Bias, standard error and 95% C.I

 is unknown and moments of T is not possible

Suppose, we find

*E*(*T*) = 

and

*Var*(*T*)= **------**→ **(1)**

As for large n  follow *N(0,1)* randomly.

Now split the data randomly in to two equal parts. Calculate estimator and  of  and estimate Bias and Variance of **(1)** by 

Let  denote the sample statistic based on a random sample of size *n*.

Denotes the sample statistics based on *n-1* values excludes.

Let

Where







where 

is known as Jack Knife estimation of the Variance of 

**exercises**

6.1 From the following population draw all possible samples of size 2 and calculate. Also find the bias in case of ratio estimate.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Population Units** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| ***YI*** | 1 | 3 | 4 | 6 | 7 | 8 | 10 | 11 |
| ***XI*** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

6.2 A sample of 34 villages was selected from a population of 170 villages to estimate the areas under Wheat in Sahiwal District during 1985. The total cultivated area under wheat during 1981 was 20,820 acres. It was found that for 1983 it was 21288. Estimate the area under wheat for 1985 using 1981 and 1983 as bench mark variable. Compare these two estimated values. Use classical ratio estimate. Compute the standard error for both the cases.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Area Under Wheat (In Acres)** | | | | | | | | |
| **S.No.** | **1981**  **(*x*1)** | **1983**  **(*x*2)** | **1985**  **(*y*)** |  | **S.No.** | **1981**  **(*x*1)** | **1983**  **(*x*2)** | **1985**  **(*y*)** |
| 1 | 401 | 70 | 50 |  | 18 | 186 | 45 | 27 |
| 2 | 630 | 163 | 149 |  | 19 | 1767 | 564 | 515 |
| 3 | 1194 | 284 | 284 |  | 20 | 604 | 238 | 249 |
| 4 | 1170 | 440 | 381 |  | 21 | 700 | 92 | 85 |
| 5 | 1065 | 250 | 278 |  | 22 | 524 | 247 | 221 |
| 6 | 827 | 125 | 111 |  | 23 | 571 | 134 | 133 |
| 7 | 1737 | 558 | 634 |  | 24 | 962 | 131 | 133 |
| 8 | 1060 | 254 | 278 |  | 25 | 407 | 129 | 103 |
| 9 | 360 | 101 | 112 |  | 26 | 715 | 190 | 175 |
| 10 | 946 | 359 | 355 |  | 27 | 845 | 363 | 335 |
| 11 | 4170 | 109 | 99 |  | 28 | 1016 | 235 | 219 |
| 12 | 1625 | 481 | 498 |  | 29 | 184 | 73 | 62 |
| 13 | 827 | 125 | 111 |  | 30 | 282 | 62 | 79 |
| 14 | 96 | 5 | 6 |  | 31 | 194 | 71 | 60 |
| 15 | 1304 | 427 | 339 |  | 32 | 439 | 137 | 100 |
| 16 | 377 | 78 | 80 |  | 33 | 854 | 196 | 141 |
| 17 | 259 | 75 | 105 |  | 34 | 820 | 255 | 263 |

6.3 From Question 6.2, estimate the area under wheat for 1985 using regression method and using 1981 as bench mark variables. Also compute standard error under both cases.

6.4 The number of cows in milk enumerated (y) from a random sample of 20 villages from a tehsil having 84 villages, as also the corresponding census figures (x) in the previous year, are given below:

|  |  |  |
| --- | --- | --- |
| **villages** | ***y*** | ***x*** |
| 1 | 237 | 155 |
| 2 | 1060 | 583 |
| 3 | 405 | 205 |
| 4 | 1085 | 738 |
| 5 | 666 | 526 |
| 6 | 542 | 284 |
| 7 | 1337 | 758 |
| 8 | 1166 | 681 |
| 9 | 399 | 143 |
| 10 | 228 | 111 |
| 11 | 813 | 616 |
| 12 | 666 | 576 |
| 13 | 681 | 540 |
| 14 | 2743 | 2242 |
| 15 | 1228 | 940 |
| 16 | 472 | 387 |
| 17 | 643 | 675 |
| 18 | 180 | 220 |
| 19 | 583 | 654 |
| 20 | 1195 | 1787 |

Given that the census estimate f the number of cows in milk in the tehsil was 74488, estimate the number of cows in milk in the current year with and without using the census information and compare the efficiencies of the estimates.

6.5 A sample survey for the study of yield of orange was conducted. From 146 villages a simple random sample of 13 village was selected. The data is given as:

|  |  |  |
| --- | --- | --- |
| **S.No. of**  **villages** | **Total No. of**  **orange trees** | **Area under orange**  **orchards (in acres)** |
| 1 | 492 | 4.80 |
| 2 | 1008 | 5.99 |
| 3 | 714 | 4.27 |
| 4 | 1265 | 8.43 |
| 5 | 1889 | 14.39 |
| 6 | 784 | 6.53 |
| 7 | 294 | 1.88 |
| 8 | 798 | 6.35 |
| 9 | 780 | 6.58 |
| 10 | 619 | 9.18 |
| 11 | 403 | 2.00 |
| 12 | 467 | 2.20 |
| 13 | 197 | 1.00 |

Given the total area under orange orchards of 146 villages is 354.78 acres, estimate the total number of orange trees in the Tehsil along with its standard error using the area under orange orchards as the auxiliary variate. Discuss the efficiency of your estimate with the one which does not make any use of the information on the auxiliary variate.

6.6 For estimating the total cattle population, a random sample of 24 villages was selected from the total 1238 villages. The number of cattle obtained in the survey is given below for each sample village, together with the corresponding census figures relating to a previous period.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No. of**  **Villages** | **Number of Cattle** | | **S.No. of**  **Villages** | **Number of Cattle** | |
| **Census** | **Survey** | **Census** | **Survey** |
| 1 | 623 | 654 | 13 | 706 | 707 |
| 2 | 690 | 696 | 14 | 1795 | 1890 |
| 3 | 534 | 530 | 15 | 1406 | 1123 |
| 4 | 293 | 315 | 16 | 118 | 115 |
| 5 | 69 | 78 | 17 | 330 | 375 |
| 6 | 842 | 640 | 18 | 218 | 212 |
| 7 | 475 | 692 | 19 | 160 | 147 |
| 8 | 371 | 292 | 20 | 210 | 297 |
| 9 | 161 | 210 | 21 | 262 | 401 |
| 10 | 298 | 555 | 22 | 262 | 401 |
| 11 | 2045 | 2110 | 23 | 185 | 199 |
| 12 | 1069 | 592 | 24 | 574 | 564 |

Compare the efficiency of the regression estimator with the ratio estimator. It is given that the number of cattle for the previous period of 1238 villages is 680,900.

6.7 An eye estimate of the fruit weight (xi) on each tree in an orchard having 150 trees was made. The total weight X was found to be 16,600 K.G. A random sample of 10 trees was taken and actual weight of fruit ‘yi’ along with eye estimate were as

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Actual weight (yi) | 67 | 58 | 71 | 39 | 41 | 42 | 39 | 51 | 46 | 61 |
| Eye estimate weight | 67 | 55 | 78 | 40 | 45 | 46 | 40 | 56 | 48 | 59 |

1. Estimate the total actual fruit weight Y by taking simple difference estimator  and find the variance of.
2. Estimate the total actual fruit weight Y by taking regression estimator  and find the variance of.

6.8 In an experimental study in a large paddy field, the weight of grain plus straw (x) and the grain yield (y ) are obtained for each of the large number of sampling units located at random over a field. The following data were obtained:



Compare the efficiencies of ratio method of estimation and that of unbiased estimators where cx, cy are the coefficient of variation of x and y respectively and cyx are coefficient of coveriation.

6.9 A sample of 34 villages were selected from a population of 140 villages for estimating the area under wheat. The following data are given:



x denotes the area under wheat for 1973 and y denotes the area for 1975. Estimate the area under wheat for 1975 by ratio estimation method and compare it with unbiased estimate. The total area under wheat cultivation for 1973 was 21288 acres.

6.10 A trained investigator makes an eye estimate of the area each parcel in a commune containing 200 parcles. This exercise produces an area of 1,160 stemmas. The areas are actually measured on the basis of a random sample of 20 parcles with the following results.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Parcel | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Estimated area | 4.8 | 5.8 | 6.0 | 5.9 | 7.6 | 6.7 | 4.7 | 5.8 | 4.4 | 5.2 |
| Actual Area | 4.5 | 5.3 | 5.8 | 6.1 | 7.1 | 6.7 | 4.2 | 5.7 | 3.9 | 5.0 |

Estimate the total area in the commune using ratio estimate, regression estimate methods and compare the efficiency of these two methods with simple method of estimation.

6.11 From artificial population of strata 2; draw all possible samples of size 2 and find 

|  |  |  |  |
| --- | --- | --- | --- |
| **Stratum 1** | | **Stratum 2** | |
| ***X*1*i*** | ***Y1i*** | ***X*2*i*** | ***Y*2*i*** |
| 1 | 3 | 5 | 6 |
| 2 | 5 | 6 | 8 |
| 3 | 6 | 7 | 9 |
| 4 | 7 | 8 | 12 |

6.12 The following data were collected in a Pilot survey to estimate the production of fresh fruits in 3 districts of Sind.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stratum No.** | **Total number of villages** | **Total area (acres)** | **Area under orchards** | **Total number**  **of trees** |
| 1 | 985 | 11253 | 10.63, 9.90, 1.45, 3.38, 5.17, 10.35 | 747, 719, 78, 201, 311, 488 |
| 2 | 2196 | 25115 | 14.66, 2.61, 4.35, 9.87, 2.42, 5.60, 4.70, 36.75 | 580, 103, 316, 739, 196, 235, 212, 1646 |
| 3 | 1020 | 18870 | 11.60, 5.29, 7.49, 7.29, 8.00, 1.20, 11.50, 1.70, 2.01, 7.96, 23.15 | 488, 227, 374, 491, 449, 50, 47, 879, 115, 115 |

Estimated total number of trees by using

1. Separate ratio estimate.
2. Combined ratio estimate.
3. Separate regression estimate.
4. Combined regression estimate.

Compare their efficiencies with stratified random sampling and simple random sampling.

6.13 If the regression of y on x is linear E(y/x) = ax + b then show that in simple random sample  give smaller variance than



6.14 Define ratio estimator for estimating the population total of character y and derive an expression for the standard error of the estimator. State the conditions under which the classical ratio estimator is best linear unbiased estimate.

6.15 If the coefficient of variation of bench mark variable x is more than twice the coefficient of variation of the variable y, then show that in large samples with simple random sampling, the classical ratio estimator  is less precise than unbiased estimator.

6.16 In sample simple random sampling without replacement the classical ratio is  obtain the exact expression for the variance of.

6.17 Values of y and x are measured for each unit in a simple random sample to estimate population ratio, which of the following estimators would you recommend to estimate *R*. (i) Always use  (ii) Always use  (iii) either use  or  depending on the conditions (given  is known) give reason for your choice.

6.18 Using the difference estimator



and regression estimator



an estimator is defined



Show that

,



and



Hence or otherwise, prove that

