Available online at www.isas.org.in/jisas
**Journal of the Indian Society of Agricultural Statistics 77(1) 2023 1–4**

# Intuition in Development of Newer Sampling Designs

**Padam Singh**

*Medanta Institute of Education and Research, Gurugram*

*Received 11 March 2023; Revised 23 March 2023; Accepted 27 March 2023*

**SUMMARY**

The paper highlights some important contributions by the author jointly with others in development of some newer sampling designs which had 'intuition' as its basis and theory derived thereafter.

*Keywords:* Inclusion probability; Inverse sampling; Odds ratio; Regression estimator; Systematic sampling; Unbiased estimator.

## 1. INTRODUCTION

'Intuition' is an instinct or gut feeling that makes one to suggest solution for a problem. In development of newer sampling designs, solution to some of the baffling problems were based on intuition. There have been some contributions by the author jointly with others in this regard based on intuitions. This paper presents details about those.

## 2. NEW APPROACH TO SYSTEMATIC SAMPLING

In systematic sampling, only the first unit is selected at random and the rest of units get automatically selected according to a pre-determined pattern.

The method of selecting a sample of size 'n' from a population of size 'N' using systematic sampling is explained as under:

Let us assume that the $N=nk$, k being an integer. Further, assume that population units N=nk are arranged in $k$ columns and $n$ rows as follows:

| **1** | **2** | **3** | **…** | | **…** | **k** |
|---|---|---|---|---|---|---|
| **k+1** | k+2 | k+3 | … | | … | 2k |
| **2k+1** | 2k+2 | 2k+3 | … | | … | 3k |
| **…** | … | … | … | r | … | … |
| **…** | … | … | … | … | … | … |
| **(n-1)k+1** | (n-1)k+2 | (n-1)k+3 | … | | … | nk |

Then, for selecting a systematic sample of n units, a random number r from 1 to $N$ is selected. Then all units of the column to which 'r' belongs are selected in the sample.

The systematic sampling because of its simplicity and operational convenience is the preferred choice for all large-scale national surveys for sampling at last stage of selection. Some of the examples are National Family Health Surveys (NFHS) and National Sample Surveys (NSS). But systematic sampling suffers from a big limitation that it is not possible to estimate the sample variance. This is because all pairs of units in the population do not have non-zero chance of selection, which is the necessary condition for unbiased variance estimation.

Dr. Padam Singh (2022) intitutively thought '***what happens when units of the row are also selected along with the units of column corresponding to the random start r'***. Surprisingly, it clicked as selecting the sample this way ensures non – zero chance of selection for all pairs of units in the sample. Based on this intuition, he proposed a new systematic sampling, the method of selection is explained as under for situations 1 & 2:

***Situation 1***

Consider a situation when $N=pq$ and $n=p+q-1$, where p and q are integers

*E-mail address*: dr.padamsingh2013@gmail.com

The selection of a sample by New Systematic Sampling comprises of the following steps:

***Step 1:*** Arrange the units of the population in *q* rows and *p* columns (*p≥q)* as under:

| | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | … | … | p |
| p+1 | p+2 | p+3 | … | … | 2p |
| … | … | … | … | … | … |
| … | … | … | … | | … |
| (q-1)p+1 | … | … | … | … | qp |

***Step 2:*** Select a random start *'r'* from 1 to *N.*

***Step 3:*** Take all units from the row and column corresponding to '*r*'.

For this sampling scheme, the $\pi_i$'s and $\pi_{ij}$'s are given by

$\pi_i = n / N$; for all, i=1,2…N

$\pi_{ij} = \frac{q}{N}$, for units in the same column,

$\pi_{ij} = \frac{p}{N}$, *f*or units in the same row and

$\pi_{ij} = \frac{2}{N}$, for rest of the units.

***Situation 2:*** **N = pq and n = q + t - 1**

In situation 1, all units of the row corresponding to random number 'r' were to be selected. But for ensuring non-zero chance of selection for every pair of units, all units from the row need not be selected, as selection of more than half of the units from the row will be adequate to ensure non-zero chance of selection for every pair of units.

In the arrangement of N = pq, let t be a number more than half of p, then

$t \geq$ p+1; if p is odd and $\geq \frac{p}{2}+1$; if p is even

2

The New Systematic Sampling in this case is as under:

Step 1: Arrange the population units in a 2-way table with q rows and p columns.

Step 2: Select a random start 'r' from 1 to N.

Step 3: Take all units from the column and t units from the row (circularly) starting with 'r'.

It has been shown that under situation 2 all pairs of units have non-zero chance of inclusion. The scheme has been extended and modified for situations when N ≠ pq.

Making use of the inclusion probabilities estimation of parameters under study could be done using standard estimation procedure.

*The selection under new systematic sampling proposed by Padam Singh (2022), tantamounts to selecting contiguous units along with selecting units with an interval. The new systematic sampling, proposed earlier by D Singh and Padam Singh (1997) exploited this feature.*

## 3. SAMPLING SCHEME FOR UNBIASED REGRESSION ESTIMATOR

When information on the average of the auxiliary variable X is available, ratio and regression estimators are used for estimation of population mean. These estimators are known to be more efficient than sample mean if the study variable is highly correlated with the auxiliary variable. But both ratio and regression estimators are biased. Importantly, availability of unit by unit information on the auxiliary variable is not required for implementation of ratio and regression estimators.

When information on the auxiliary variable is available for every unit of the population then Midzuno and Sen (1952) suggested a method of sampling in which first unit is selected by PPS and remaining (n-1) units by SRS WOR. For this sampling scheme the probability of selecting the sample was proportional to sample mean of the auxiliary variable. Incidentally, for this sampling scheme it was seen that the ratio estimator becomes unbiased.

In a search for unbiased regression estimator, Singh and Srivastava (1980) proposed a sampling scheme for which the usual regression estimator becomes unbiased. The intuition for proposing this scheme was 'what happens if sample is selected by probability proportional to sample variance of the auxiliary variable'. This intuition worked as for such a sampling scheme, the regression estimator became unbiased.

When information on the auxiliary variable is available for every unit of the population, the Sampling scheme ensuring probability proportional to sample

variance of the auxiliary variable proposed by Singh and Srivastava (1980) consists of the following steps:

Step 1: Select 2 units, say i, j with their probability of joint selection proportional to $(X_i - X_j)^2$

Step 2: Select remaining (n – 2) units from the remaining units of the population by SRSWOR.

For this sampling scheme, the probability of selecting the sample is given by

$$P_s = s_x^2 / (NCn) S_x^2$$

It was shown that under this sampling scheme the usual regression estimator

$t_s = \bar{y} + b(\bar{X} - \bar{x})$ is unbiased for $\bar{Y}$.

If x and y follow a bivariate normal distribution, the variance of proposed estimator reduces to the variance of usual regression estimator.

The performance of regression estimator under proposed sampling scheme was seen as highly satisfactory, particularly when there is large departure of line of regression of y on x from origin, that is the situation where usual regression estimator is preferred.

## 4. UNBIASED ESTIMATION OF ODDS RATIO

For a usual 2X2 contingency table of *antecedent* versus *outcome*, in case control studies, those with positive outcome are generally referred to as cases and with negative outcome as controls. Similarly, those with antecedent present are termed as exposed and antecedent absent as unexposed.

In such studies a population of N units can be considered as consisting of $N_1$cases (with disease) and the remainder $N_2 = (N - N_1)$ as controls (free from the disease). Further suppose A out of $N_1$ and B out of $N_2$ are exposed to an environment, represented as follows:

| | **Cases (with disease)** | **Controls (free from the disease)** | **Total** |
|---|---|---|---|
| **Exposed** | A | B | A + B |
| **Unexposed** | C = ($N_1$ – A) | D = ($N_2$ – C) | C + D |
| **Total** | $N_1$ = A+C | $N_2$ = B+D | A+B+C+D = N |

The parameter of interest in case control studies is the Odds Ratio, which is ratio of odds of exposed among cases to that among controls given by

$\theta = \frac{AD}{BC}$, which approximates relative risk for low prevalence diseases

For estimation of odds ratio, samples of cases (i.e., diseased persons) and controls (i.e., non-diseased persons) are selected independently and the number of exposed in each is recorded.

Thus essentially samples of $n_1$ from cases and $n_2$ from controls are from populations of (A + C) and (B + D) respectively. Suppose in a sample of $n_1$ cases 'a' are observed to be exposed to the environment, and in the sample of $n_2$ controls 'b' are observed as exposed, then we have the following.

| | **Cases (with disease)** | **Controls (free from the disease)** |
|---|---|---|
| **Exposed** | a | b |
| **Unexposed** | c (= $n_1$ - a) | d (= $n_2$ – b) |
| **Total** | $n_1$ = a + c | $n_2$= b + d |

In such case control studies the total sample sizes ($a + c = n_1$) and ($b + d = n_2$), are fixed but a, b, c, and d are random variables.

Cornfield (1951) proposed the estimator of odds ratio $(\theta)$ given by

$t = \frac{ad}{bc}$, sample odds ratio

Evidently, above estimator is biased.

If the risk factor is strongly associated with the disease, then the frequencies b and c are generally very small or may even be zero. The examples of this are lung cancer and cigarette smoking, oral cancer and tobacco chewing, etc. In situations when either b or c is zero, the estimator becomes infinity. Teststatistic 't' for testing the significance of odds ratio becomes indeterminate. Thus, it is desirable to modify the sampling scheme so that the observed frequencies in the cells, "not exposed but diseased", and "exposed but not diseased", are fixed to some minimum number. *To circumvent this problem, Padam Singh and Abha Aggarwal (1991) intuitively considered using inverse sampling in case control studies to ensure such a minimum.*

The proposed scheme using inverse sampling under case control studies is as follows:

Select a sample of cases by SRSWOR, until c′ unexposed individuals are observed. In this situation the total sample size of cases will be a random variable.

Similarly, select a sample of controls by SRSWOR until b′ exposed individuals are observed. Here also the total sample size of controls will be a random variable.

Suppose in this way we obtain a sample represented as follows:

| | **Cases (with disease)** | **Controls (free from the disease)** |
|---|---|---|
| **Exposed** | a′ | b′ |
| **Unexposed** | c′ | d′ |
| **Total** | a′ + c′ | b′ + d′ |

In this scheme b′ and c′ are fixed but a′ and d′ & consequently a' + c' and b' + d' are random variables.

For this sampling scheme the usual odds ratio can be considered as an estimator of Θ, given by

$$t_1 = \frac{a'd'}{b'c'}$$

In the situation under study b' and c' are rare and N is large. The underlying distribution for rare attribute is hypergeometric. It is well known that the hypergeometric distribution tends to negative binomial distribution when 'N' is large.

As under inverse sampling a' and d' follow the negative binomial distribution, surprisingly sample odds ratio became unbiased for $\theta$ = AD/BC.

This is explained as under

$$E(a') = \frac{c'\tau_1}{\mu_1} \text{ where } \tau_1 = \frac{A}{\overline{A}+\overline{C}},\ \mu_1 = 1-\tau_1$$

and

$$E(d') = \frac{b'\tau_2}{\mu_2} \text{ where } \tau_2 = \frac{B}{\overline{B}+\overline{D}},\ \mu_2 = 1-\tau_2$$

Since a' and d' are independent,

$$E(a'd') = E(a').E(d') = \frac{c'\tau_1}{\mu_1}\frac{b'\tau_2}{\mu_2}$$

Thus, $E(t_1) = \frac{AD}{BC}$

Use of inverse sampling in case control studies not only provided a solution to the problem of ensuring minimum b and c but also ensured unbiased estimation of parameter $\theta$.

On comparing the efficiency, it was seen that the proposed estimator performed better in situations where the relative risk is high.

## REFERENCES


Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of lung, breast and cervix. *Journal of the National Cancer Institute*. 1269-75.

Midzuno, H. (1952). On the Sampling System with Probability Proportional to Sum of Sizes. *Annals of the Institute of Statistical Mathematics*, **3**, 99-107.

Sen A.R. (1952). Present status of probability sampling and its use in estimation of farm characteristics. *Econometrica*. **20**, 103.

Singh, D. and Singh, Padam (1977). New systematic sampling, J*ournal of statistical planning and inference*, 163-177.

Singh, Padam and Srivastava, A.K. (1980). Sampling schemes providing unbiased regression estimators. *Biometrika*. **67**(1), 205-9.

Singh, Padam and Aggarwal, A.R. (1991). Inverse sampling in case control studies. *Environmetrics*. **2**(3), 293-99.

Singh, Padam. (2022). New systematic sampling–II. *Journal of Indian Society of Agricultural Statistics*. **76**(2), 47-58.