Class Notes on Sample Survey

Cheenta School of Statistics and Data Science

March 24, 2023

Contents

Abstract

These are informal class notes on the topic Sample Survey, which will be covered very briefly in our AIMStat course. It may not be followed exactly in class, but we encourage you to go through it. We sincerely hope that these notes along with attending the classes will make Sample Survey less boring for you. An important point to be mentioned is that, we will be doing only and only the basics of Sample Survey which will include a revision of B1 level Probability Theory too. Hence, don't be shocked to see probability problems emerging out of nowhere in the notes. If you follow the classes, the style will be familiar to you and hence you will find it easier to read.

1 Introduction

In statistical inference, whatever problems we have encountered can be broadly categorized into two parts, namely *Estimation* and *Testing of Hypothesis*. However, if you think about solving problems in those contexts, a random sample from the population $X_1, X_2, ..., X_n$ was always given to you. Sample Survey is the branch of Statistics which deals with the question *how to draw a good sample from a population?*. Since, the accuracy of our results obtained does not only depend on how good our methods are, but also the quality of the data we have at hand, Sample Survey is a topic which we just cannot ignore. If you work with real life data in future, you will be able to understand the importance of having a "good" sample, without which your conventional estimates or testing procedures can fail miserably. Thus, throughout this course, our objective will be to draw a sample which is a "good" representative of the population we are interested in. As we define everything mathematically, it will be easy to see that the basics are nothing but a beautiful application of discrete probability.

Probability Problem 1 : The numbers a and b are chosen without replacement from the set $1, 2, .., N$. Find $P(a > b)$. If you select three numbers, find the probability that the second number lies between the first and the third number.
Solution : To be done in class.

2 Simple Random Sampling

Consider a population with N number of units denotes by $U_1, U_2, ..., U_N$. Let, $\mathbf{y}$ be our variable of interest, also known as the *study variable*. Suppose Y_i denote the value of the study variable for the i^{th} population unit, $i = 1, 2, ..., N$.

Example 1. Our population can be a class of 200 students. Clearly, here $N = 200$. Say we are interested in estimating the average height of those students, then **height** is our variable of interest $\mathbf{y}$. According to our notation, Y_i is the *height of the* i^{th} *student*, $i = 1, 2, ..., 200$.

Define the following quantities :

- $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ (Population Mean)
- $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \bar{Y})^2$ (Population Variance)
- $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i - \bar{Y})^2$ (We will see the convenience of this notation at many instances)

Our goal is to **estimate some population parameter**. It can be the mean of the population, or the median, the standard deviation, co efficients of skewness, kurtosis anything. In *Example* 1, the parameter we wanted to estimate was the population mean.

We draw a random sample of size n from the population. This kind of sampling scheme where each of the population units have the same probability of being included in the sample is known as Simple Random Sampling (SRS). *Sample is defined to be an ordered collection of units from the population.* The sample can be drawn in two ways, *with replacement* (WR) or *without replacement* (WOR). Suppose y_i denote the value of the study variable for the i^{th} sampled unit, $i = 1, 2, ..., n$.
Define the following quantities :

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (Sample Mean)
- $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ (Sample Variance)

Remark : In case of SRSWR, total number of possible samples of size n from a population of size N is N^n. In case of SRSWOR, that is ${}^{N}P_n$.

2.1 Inclusion Probabilities

Inclusion probability of the i^{th} $(i = 1, 2, .., N)$ population unit U_i is defined as the probability of that unit belonging to a particular sample of size n. It is denoted by π_i. This is also known as first order inclusion probability.

$$\pi_i = P(i^{th} \text{ unit is included in a particular sample of size } n) = P(U_i \in \text{sample})$$

Similarly, the second order inclusion probability is defined as

$$\pi_{ij} = P(\text{the } i^{th} \text{ and the } j^{th} \text{ population unit belongs to a particular sample}) = P(\{U_i, U_j\} \in \text{sample}).$$

Exercise 1 : Show that, $\pi_i = 1 - (\frac{N-1}{N})^n$ for SRSWR and $\pi_i = \frac{n}{N}$ for SRSWOR.
Solution : To be done in class.

Exercise 2 : Find the expressions for π_{ij} for both SRSWR and SRSWOR.
Solution : To be done in class.

Probability Problem 2 : A sample of size n is drawn from a population consisting of N units with replacement. If m be the number of distinct units in the sample, find $E(m)$ and $V(m)$.
Solution : To be done in class.

Probability Problem 3 : From an urn containing N tickets numbered $1, 2, ..., N$, we draw n tickets one by one at random. What is the probability that the highest number drawn is equal to M $(1 \leqslant M \leqslant N)$ if the draws are made (a) with replacement? (b) without replacement?
Solution : To be done in class.

2.2 Estimation of the Population Mean

The very common problem faced often by us is to get an idea of the population mean / average value of a particular characteristic of the population. Thus, this estimation problem is of special interest in sample survey. We will find an unbiased estimator (which is considered as a so called "desirable" property of an estimator) of the population mean and also try to estimate it's standard error. Intuitively, the sample mean should be a "good" estimator of the population mean. Let us see if our intuition matches with mathematical derivations. We will be as consistent as possible with our notations and terminologies defined earlier.

Let, $Y_1, Y_2, ..., Y_N$ denote the population units and $y_1, y_2, ...y_n$ denote the sampled units.For simple random sampling with replacement (SRSWR), note that $\forall i = 1, 2, ..., n$, we have

$$P(y_i = Y_j) \ = \ \frac{1}{N} \ \forall j = 1, 2, .., N$$

i.e. each population unit has equal probability of being selected at each draw.

$$E(y_i) \ = \ \sum_{j=1}^{N} Y_j P(y_i = Y_j) \ = \ \sum_{j=1}^{N} Y_j \ \frac{1}{N} \ = \ \frac{1}{N} \sum_{j=1}^{N} Y_j \ = \ \bar{Y} \ \forall i = 1, 2, ..., n$$

$$\implies E(\bar{y}) = E(\frac{1}{n}(y_1 + y_2 + ... + y_n)) \ = \ \frac{1}{n} \sum_{i=1}^{n} \bar{Y} \ = \ \bar{Y}$$

Hence, sample mean is an unbiased estimator of the population mean in SRSWR. Whenever we suggest an estimator for a population parameter, as a statistician we must report the accuracy (which is given by bias) and precision (given by variance) of that estimator.

$$V(y_i) \ = \ E(y_i^2) - (E(y_i))^2 \ = \ \sum_{j=1}^{N} Y_j^2 P(y_i = Y_j) \ - \bar{Y}^2 \ = \ \sum_{j=1}^{N} Y_j^2 \ \frac{1}{N} \ - \bar{Y}^2 \ = \ \frac{1}{N} \sum_{j=1}^{N} (Y_i - \bar{Y})^2 \ = \ \sigma_y^2$$

$$\implies V(\bar{y}) \ = \ V(\frac{1}{n} \sum_{i=1}^{n} y_i) \ = \ \frac{1}{n^2} \sum_{i=1}^{n} V(y_i) \ = \ \frac{1}{n^2} \sum_{i=1}^{n} \sigma_y^2 \ = \ \frac{\sigma_y^2}{n}$$

To give an unbiased estimator of σ^2, we can start by taking the expectation of the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

$E((n-1)s^2) = E(\sum_{i=1}^{n} (y_i - \bar{y})^2) = E(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2) = E(\sum_{i=1}^{n} y_i^2) - nE(\bar{y}^2) = \sum_{i=1}^{n} (\sigma_y^2 + \bar{Y}^2) - n(\frac{\sigma_y^2}{n} + \bar{Y}^2) = (n-1)\sigma_y^2$
$\implies E(s^2) = E(\sigma_y^2) \implies E(\frac{s^2}{n}) = E(\frac{\sigma^2}{n}) = V(\bar{y})$

Exercise 3 : Suggest an uniased estimator for $\bar{Y}$ in case of SRSWOR. Find the variance of that estimator and give an unbiased estimator of that variance.
Solution : Check that

$$E(\bar{y}) = \bar{Y} \text{ and } V(\bar{y}) = (\frac{1}{n} - \frac{1}{N}) S_y^2$$

Exercise 4 : Suppose you have bought a pack of N (unknown) cards. You have a little annoying sister/ brother who

- Draws a card
- Notes down the number on it
- Puts it back
- Repeats the task

In this way he/ she has got n numbers i.e has drawn n times. Then he/she wants to embarrass you in front of your parents and asks you "Hey, you are studying Statistics since 3 years, can you tell me how many cards are actually in the pack?". Clearly, your task is to find an unbiased estimator of N. Also, find the variance of your estimator to get a feel of winning a cold war against your sibling.
Solution : To be done in class.

2.3 Survey Sampling and Discrete Probability Distributions

Suppose a bag contains M red balls and $N-M$ blue balls. We select n balls out of the bag with replacement. Clearly, each draw is a Bernoulli trial with probability of success (getting a red ball) $\frac{M}{N}$. Let X be the random variable denoting the number of red balls drawn. Then

$$P(X = x) = \binom{n}{k} (\tfrac{M}{N})^x (1 - \tfrac{M}{N})^{n-x} \; \forall \; x = 0, 1, 2, .., n$$

Thus, $X \sim Bin(n, \frac{M}{N})$. This shows the relation between the Binomial distribution and SRSWR.
Similarly, if we select n balls out of the bag without replacement, it is easy to see that X follows a Hypergeometric distribution with parameters (N, M, n).

Exercise 5 : Let X follow a Hypergeometric distribution with parameters (M,N,n). Find the distribution of X as $N \to \infty$ and $\frac{M}{N} \to p$. What is the interpretation of the result you get?
Solution : Homework.

Exercise 6 : Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ independently. Find the distribution of $X|X + Y = k$. What is the interpretation of the result you get?
Solution : Homework.

Probability Problem 4 : You go to play cricket with your friends daily. One day, during the toss, someone argues that the coin is biased with $P(H) = p$ (unknown). Your friends ask you to provide a solution to this problem as you know probability better than them. How will you tackle this situation? Clearly, your task is to simulate an unbiased coin from a biased coin. **Solution :** To be done in class.

Probability Problem 5 : For any two events A and B, show that

$$P(A \cap B)^2 + P(A \cap B^c)^2 + P(A^c \cap B)^2 + P(A^c \cap B^c)^2 \geqslant \frac{1}{4}$$

Solution : To be done in class.

3 Stratified Random Sampling

By this time it should be familiar to us that we want to design our sampling scheme in such a way that the sample is a good representative of the population. Thus, if we think a bit deeply, in simple random sampling, we are implicitly assuming that the population is homogeneous with respect to the feature / characteristic under study. But, this may not be the case always. In such situations, we need to modify our sampling schemes accordingly. Stratified Random Sampling is one of the simplest modifications over SRS which is adopted when the population can be divided into several groups with respect to the variable of our interest. Suppose if we think of our earlier example of "height of an individual" being the variable of our interest, then clearly the entire population of a particular state (say) is not homogeneous with respect to height. We can roughly divide the entire population into some sub populations : 0-2 years, 3-5 years, 5-7 years, 7-9 years, 9-11 years, 12-15 years, 16-18 years, 18-21 years, 21 and above, say. then each of these sub populations are more or less homogeneous with respect to height. The main idea of stratified sampling is to draw different samples from different sub populations and then merge them instead of drawing one single sample. Each of these sub populations is known as a *Stratum (plural is Strata)*. **Detailed explanations will be discussed in class**.

3.1 Sampling Scheme

Say we want a sample of size n from a population of size N. Suppose the population is divided into 5 strata of sizes N_1, N_2, N_3, N_4, N_5 respectively. We draw samples of size n_h from the $h^{th}, h = 1, 2, 3, 4, 5$ strata consisting of N_h units subject to the constraint $\sum_{i=1}^{5} n_i = n$. Define $W_h = \frac{N_h}{N}$. (These can be thought of as weights, in terms of size of the strata)

Let Y_{hi} denote the value of the study variable for the i^{th} unit of the h^{th} stratum. For the h^{th} stratum, define

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}, \quad \sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2, \quad S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$$

Here also, we will focus on estimating the population mean $\bar{Y}$. Let us see how.

Probability Problem 6 : Consider n coins each with $P(H) = p$. Coins are tossed simultaneously and then coins which show head are tossed again. If p' is the probability of getting a head in the second round of tossing, find the probability distribution of the number of heads obtained in the second round of tosses.
Solution : To be done in class.

3.2 Estimation of the Population Mean

Before going for estimation, let us note that

$$\sum_{h=1}^{5} W_h \bar{Y}_h = \sum_{h=1}^{5} \frac{N_h}{N} \bar{Y}_h = \frac{1}{N} \sum_{i=1}^{N} Y_i = \bar{Y}$$

Thus, we can just find an unbiased estimator for each of the $\bar{Y}_h$ and then use linearity of expectation to get our required estimator. Let y_{hi} denote the value of the study variable for the i^{th} sampled unit from the h^{th} stratum. For the h^{th} stratum, define

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}, \quad s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

Then, clearly by the principles of simple random sampling we have $E(\bar{y}_h) = \bar{Y}_h$ which by linearity of expectation gives us $E(\sum_{h=1}^{5} W_h \, \bar{y}_h) = \bar{Y}$. Hence, $\bar{y}_{st} = \sum_{h=1}^{5} W_h \, \bar{y}_h$ is our unbiased estimator in this case. Now again, whenever we give an estimator, we should give its standard error too. Thus, we need to find the variance of $\bar{y}_{st}$ at the first place.

Exercise 7 : Find the variance of $\bar{y}_{st}$ if the stratified sampling is done (i) with replacement (ii) without replacement.
Solution : Homework.

Exercise 8 : Consider a collection of N cards numbered $1, 2, ..., N$ where $N \geqslant 2$. A card is drawn at random and set aside. Suppose that n cards are selected from the remaining $N-1$ cards using SRSWR and the numbers are noted as $Y_1, Y_2, ..., Y_n$. if $S = \sum_{i=1}^{n} Y_i$ then find $E(S)$ and $Var(S)$.
Solution : To be done in class.

Probability Problem 7 : Two policemen are sent to watch a road that is 1 km long. Each of the two policemen is assigned a position on the road which is chosen according to a uniform distribution along the length of the road and independent of the other's position. Find the probability that the two policemen will be less than $\frac{1}{4}$ km apart when they reach their assigned positions.
Solution : To be done in class.

Exercise 9 : Consider a stratified random sampling scheme with replacement with 2 strata. Let N_i and σ_i^2 respectively denote the size and the variance of the i^{th} stratum $\forall i = 1, 2$. Define $\lambda = \frac{N_1}{N_2}$ and $d = \frac{\sigma_1}{\sigma_2}$. If V_0 is the variance of the usual unbiased estimator of $\bar{Y}$ for the *optimum* choice of n_1, n_2 and V_e is the variance of the same estimator for the choice $n_1 = n_2$, then show that

$$\frac{V_e - V_0}{V_0} = \left(\frac{1 - \lambda d}{1 + \lambda d}\right)^2$$

Solution : To be done in class.

Exercise 10 : An SRSWOR of size 2 is drawn from a population containing 3 units U_1, U_2, U_3. Consider an estimator T of $\bar{Y}$ as follows

$$\begin{aligned} T &= \tfrac{1}{2}(Y_1 + Y_2) \text{ if } U_1, U_2 \text{ are in the sample} \\ &= \tfrac{Y_1}{2} + \tfrac{2Y_3}{3} \text{ if } U_1, U_3 \text{ are in the sample} \\ &= \tfrac{Y_2}{2} + \tfrac{Y_3}{3} \text{ if } U_2, U_3 \text{ are in the sample} \end{aligned}$$

Show that T is an UE of $\bar{Y}$. Compute the variance of T and comment.
Solution : Homework.

4 Cluster and Two Stage Sampling

4.1 Sampling Scheme

Sometimes a population is naturally divided into or can be conveniently thought to be divided into a number of clusters or groups of units. Then, the structure of the population is as follows :

Population Units	Clusters	Size
$U_{11}, U_{12}, ..., U_{1M_1}$	U_{1*}	M_1
$U_{21}, U_{22}, ..., U_{2M_2}$	U_{2*}	M_2
.....	....	....
.....	....	....
$U_{N1}, U_{N2}, ..., U_{NM_N}$	U_{N*}	M_N

Total number of units in the population $= \sum_{i=1}^{N} M_i$.
Let U_{ij} be the j^{th} unit in the i^{th} cluster ($j = 1, 2, .., M_i,\ i = 1, 2, .., N$).

These clusters are called first stage units (FSU). In such situations, instead of drawing a sample directly from the population, it is convenient and cheaper to select a sample of FSUs and then

1. Either enumerate all the selected FSUs completely OR
2. Select samples from the selected FSUs

The first sampling scheme is known as **Cluster Sampling** and the second sampling scheme is known as **Two Stage Sampling**.

4.2 Estimation of the Population Mean

Consider for simplicity $M_1 = M_2 = ... = M_N = M$. Suppose n first stage units (FSU) are selected by SRSWOR and from each of the selected FSUs, m second stage units (SSU) are selected again by SRSWOR. Now, we define the following quantities.

$\bar{\bar{Y}} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij} = \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_i$, $S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2$, $S_{2i}^2 = \frac{1}{M-1}\sum_{j=1}^{M}(Y_{ij} - \bar{Y}_i)^2$, $S_2^2 = \frac{1}{N}\sum_{i=1}^{N} S_{2i}^2$

$$\bar{\bar{y}} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij},\ \ s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2,\ \ s_{2i}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2,\ \ s_2^2 = \frac{1}{n}\sum_{i=1}^{n} s_{2i}^2$$

$$f_1 = \frac{n}{N},\ f_2 = \frac{m}{M}$$

Result 1: $E(\bar{\bar{y}}) = \bar{\bar{Y}}$ and $V(\bar{\bar{y}}) = \frac{1-f_1}{n}S_1^2 + \frac{1-f_2}{mn}S_2^2$

Result 2: $E(\frac{1-f_1}{n}s_1^2 + \frac{f_1(1-f_2)}{mn}s_2^2) = V(\bar{\bar{y}})$

Exercise 11: Prove the two results mentioned above.
Solution: Homework.

Exercise 12: A chocolate frog costs 5 sickels. Each of them contain exactly one card with a moving picture of any of the four founders of Hogwarts : Godric Gryffindor, Rowena Ravenclaw, Helga Hufflepuff and Salazar Slytherin i.e. there are 4 types of cards. You go on buying chocolate frogs till you get all 4 types of cards in your collection. Calculate your expected expenditure.
Solution: To be done in class.

Exercise 13: An SRSWOR of size $n(n_1 + n_2)$ with mean $\bar{y}$ is drawn from a finite population of size N and an SRSWOR of size n_1 is drawn from it with mean $\bar{y}_1$. Show that

1. Cov$(\bar{y}, \bar{y}_1 - \bar{y})$=0
2. $V(\bar{y}_1 - \bar{y}) = S^2(\frac{1}{n_1} - \frac{1}{n})$
3. $V(\bar{y}_1 - \bar{y}_2) = S^2(\frac{1}{n_1} + \frac{1}{n_2})$
4. Cov$(\bar{y}_1, \bar{y}_2)$= $-\frac{S^2}{N}$

Solution: To be done in class.

5 Miscellaneous Exercises

1) 3 independent samples of sizes n_1, n_2, n_3 are respectively drawn from a finite population of size N by SRSWR sampling scheme such that $n_1 + n_2 + n_3 = n$. If p_1, p_2, p_3 denote the sample proportions of an attribute A, obtain the combined estimate of the population proportion P_A and an expression for the relative standard error (with respect to the mean) of the error.

2) Determine the sample size required to estimate the population proportion using sample proportion based on an SRSWR sample having a maximum absolute error of 5% with confidence probability at least 0.95. (Use large sample approximation)

3) Let n be the required sample size to estimate the proportion of workers in a finite population with coefficient of variation α % in SRSWR. Find the sample size n^* to estimate the proportion of non workers with the same precision.

4) An SRSWR of size 3 in drawn from a population of size N. Find the probabilities for the sample to have 1,2 and 3 distinct units. Hence, show that the sample mean $\bar{y}^*$ based on only distinct units of the sample is unbiased for the population mean. Also, find $V(\bar{y}^*)$ and show that $V(\bar{y}^*) \leqslant V(\bar{y})$.

5) A sample of size 4 is drawn from a population of size 8. Suppose, units U_1 and U_8 are included in every sample and SRSWOR of size 2 is drawn from the units $U_2, ..., U_7$. Show that $T = \frac{1}{8}(Y_1 + Y_8 + 6\bar{y}^*)$ is an UE of the population mean where $\bar{y}^*$ is the sample mean of the two units drawn. Also find the variance of this estimator.

Probably Cochran

6) In a particular population of size N, the variate value of one of the units is known to be Y_1. An SRSWOR of size n is drawn from the remaining $N - 1$ units and the sample mean $\bar{y}^*$ is calculated. Consider an estimator T of the population total as $T = Y_1 + (N - 1)\bar{y}^*$. Show that T is an UE for Y_{TOTAL} and has a smaller variance than the usual unbiased estimator $N\bar{y}$.

Probably Desraj

References

Chaudhuri, Arijit. *Essentials Of Survey Sampling. India: Prentice-Hall Of India Pvt. Limited, 2010*

Sampath, S. *Sampling theory and methods. India: CRC Press, 2001*

Rice, John A. *Mathematical Statistics and Data Analysis. Austria: Cengage Learning, 2007*