Theory of Survey Sampling

Soumyadeep Das

Department of Statistics, Bidhannagar Government College, Kolkata, India

1 Introduction

Stages of Statistical study:

1. Collection of data
2. Scrutiny of data
3. Condensation of data
4. Analysis of data
5. Interpretation/Inference/Conclusion

For a plausible/reasonable conclusion:

- Sample should be good i.e. a proper representation of the population.
- Decision rule should be good.

So you need to know what is population and what is a sample and how a sample can be drawn from a population in different scenarios. This story will be discussed in this course. Moreover, you will learn how to make a good decision in your 'statistical inference' course.

There are two main procedures of data collection. One is complete enumeration and another is sampling. There are several advantages of sampling over complete enumeration. Principal steps of a sampling, questionnaire, non responses, sources of error etc. are related to complete enumeration as well as sampling procedure. You have to study the details from Fundamental Vol-II.

Two types of sampling:

1. Random sampling: A random or probability sample is a sample drawn in such a manner that each population unit has a predetermined probability of selection.

2. Non-random sampling: Judgment/purposeful/quota sampling.

Before going into the details, we need some concepts, definitions and notations. Lets discuss these first.

1. N = Population size (known).

2. $\boldsymbol{U} = \{U_1, U_2, \ldots, U_N\}$ is a survey population is the collection of all units of a specific type in a given region at a particular period of time. E.g.- population of India, all the fishes in a given pond, all the Covid-19 patients in the world, all the trees in a given forest, etc.

3. $U_j = j$ th identifiable population unit on which observations can be made or from which the required statistical information can be ascertained according to a well-defined procedure. E.g.- person, farm, factory, etc.

4. **Reporting unit:** The unit which actually supplies the required statistical information. E.g.- Head of a family for any family budget enquiry, HOD of a department for students' evaluation survey, etc.

5. **Parameter:** Any function of the values of all population units. Let Y_j is the associated value for U_j i.e. $Y_j = Y(U_j)$ where Y is the variable under study. We denoting $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)$, population total $Y = \sum_{i=1}^{N} Y_i$ or population mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ are the parameters.

6. **Sampling unit:** Elementary units or groups of such units, which, besides being clearly defined, identifiable and observable, are convenient for purpose of sampling.

7. **Sampling frame:** For using sampling methods in the collection of data, it is essential to fave a frame of all the sampling units belonging to the population to

be studied with the proper identification particulars and such a frame is termed as 'sampling frame'.

8. **Sample:** $\boldsymbol{s} = \{u_1, u_2, \ldots, u_n\}$ of size n is a sequence of ordered elements of $\boldsymbol{U}$. They may or may not be unique. The set of all possible sequences $\boldsymbol{s}$ is denoted by $\mathscr{S}$ and is called the **sample space**. Moreover, we write $y_i = Y(u_i)$ for $1, 2, \ldots, n$.

9. **Selection Probability:** Probability of selecting a sample $\boldsymbol{s}$ is denoted as $p(\boldsymbol{s})$ with $\sum_{\boldsymbol{s}\in\mathscr{S}} p(\boldsymbol{s}) = 1$.

10. **Sampling design:** $(\boldsymbol{s}, p(\boldsymbol{s}), \boldsymbol{s} \in \mathscr{S})$.

11. **Sampled unit:** The sampling units selected in the sample may be termed as sampled units.

12. **Statistic:** $t = t(\boldsymbol{s}, \boldsymbol{Y})$ with the values of Y_i for only $i \in \boldsymbol{s}$ but are free of every Y_i, $i \notin \boldsymbol{s}$, is called a statistic. It may be used to estimate a population parameter.

13. **Inclusion Probability:**
    a) Of order 1: $\pi_i = P(U_i \in \boldsymbol{s}) = \sum_{\boldsymbol{s}\ni i} p(\boldsymbol{s})$
    b) Of order 2: $\pi_{ij} = P(U_i, U_j \in \boldsymbol{s}) = \sum_{\boldsymbol{s}\ni i,j} p(\boldsymbol{s})$

14. **Expectation of a statistic:** $E(t) = \sum_{\boldsymbol{s}\in\mathscr{S}} p(\boldsymbol{s}) t(\boldsymbol{s}, \boldsymbol{Y})$.

15. **Unbiasedness:** A statistic $t(\boldsymbol{s}, \boldsymbol{Y})$ is said to be **design unbiased** for a parameter θ if $E(t) = \theta$, for all $\boldsymbol{Y} \in \mathbb{R}^N$

References

1. Goon AM, Gupta MK and Dasgupta B (2008): Fundamentals of Statistics, Vol-II.

2. Cochran WG (1984): Sampling Techniques.

3. Des Raj and Chandhok P (1998): Sample Survey Theory.