STA304 - Surveys, Sampling and Observational Data

Heather Tan

Jan - Apr 2020

Contents

1 Basic Information **2**
1.1 Random Variable . . . 2
1.2 Statistics Info . . . 2
1.2.1 Population Total . . . 2
1.2.2 Population Proportion . . . 2
1.2.3 Population Ratio . . . 3

2 Basic idea of sample survey design **3**
2.1 Basic Notations from statistics . . . 3
2.2 Some technical definitions . . . 3

3 Design of Sample Survey **4**
3.1 General Procedure: . . . 4
3.2 How to select a sample . . . 4
3.2.1 Cencus . . . 4
3.2.2 Probability, or random sampling . . . 4
3.2.3 Nonrandom/nonprobability sampling . . . 5
3.3 Other methods of Data Collection . . . 6
3.3.1 Sources of Errors in Surveys . . . 6
3.3.2 Methods of data collection . . . 7
3.4 Sampling design process . . . 7

4 Simple Random Sampling **7**
4.1 Simple Random Sample without replacement . . . 8
4.1.1 Table of random numbers . . . 8

5 Inference of SRS **8**

5.1 Probability . . . 8

5.1.1 SRS . . . 8

5.1.2 SRSWR . . . 8

5.2 Inclusion probability of SRS . . . 9

5.2.1 Covariance under SRS . . . 9

5.3 Inference of Sample mean μ . . . 9

5.3.1 $\hat{\mu}$ is unbiased: . . . 9

5.3.2 Variance of $\hat{\mu}$ under SRSWR . . . 10

5.3.3 Variance of $\hat{\mu}$ under SRS . . . 10

5.3.4 Summary of population mean $\mu = \bar{y}$ . . . 10

5.4 Inference on population variance σ^2 . . . 10

5.4.1 Estimation of σ^2 Under SRSWR . . . 11

5.4.2 Estimation of σ^2 Under SRS . . . 11

5.4.3 Summary of Unbiased Estimation of σ^2 . . . 11

5.5 Inference of Population Total τ . . . 11

5.6 Inference of Population proportion p . . . 11

1 Basic Information

Definition 1.1. A Random Experiment is the process of observing the outcome of a chance event

Definition 1.2. The elementary outcomes are all possible results of the random element

Definition 1.3. The sample Space (Ω) is the set or collection of all the elementary outcomes.

E.g if the event was a coin toss

- The random experiment consist of recording its outcome
- the elementary outcomes are heads(H) and tails(T)
- $\Omega = \{H, T\}$

1.1 Random Variable

A random variable Y is a real-valued function defined over a sample space. It can be sued to identify numerical events that are of interest in an experiment.

Definition 1.4. A random variable Y is said to be discrete if it can assume only finite or countable infinite number of distinct values

Definition 1.5. A random variable Y is said to be continuous if it can take on any value of an interval.

1.2 Statistics Info

1.2.1 Population Total

$$\tau_y = \sum_{i=1}^{N} y(e_i) = \sum_{i=1}^{N} y_i = \sum_{i=1}^{k} N_i y_i$$
$$\tau_y = N\mu_y, \mu_y = \frac{1}{N}\tau_y$$

1.2.2 Population Proportion

Population Proportion Proportion of elements which poses certain property, or belong to certain specified group. Define variable y taking two values:

$$y(e) = \begin{cases} 0 & \text{e does not have the property} \\ 1 & \text{e has the property} \end{cases}$$

$$p = \frac{1}{N}\sum_{i=1}^{N} y(e_i) = \frac{M}{N} = \frac{\text{Number of elements with the property}}{Totalnumberofelements} = \mu_y$$

If two variables x,y are considered

1.2.3 Population Ratio

When two variables are considered, **Ratio of their means, or their totals:**

$$R = \frac{\mu_u}{\mu_x} = \frac{N \cdot \tau_y}{N \cdot \tau_x} = \frac{\tau_y}{\tau_x} = R_{y/x}$$

2 Basic idea of sample survey design

- **Sample Survey** is a partial investigation of the finite population using samples. The purpose of sample survey is to obtain information about the population.
- **Population** is a group of units defined according to the aims and objects of the survey.
- **Sampling** is the selection of part of the population.
- **Sampling Method** is a scientific and objective procedure of selecting units from a population. It provides a sample that is expected to be representative of the population as a whole. It also provides procedures for estimation of the population parameters.

2.1 Basic Notations from statistics

- Population: $E = \{e_1, \cdots, e_N\}$
- Population size: N
- elements e, e_i
- Variable: y,x z,t $\cdots$ $e \implies y(e)$
- Range: $\{y(e), e \in E\}$
- **Distribution, frequency distribution**: Proportion/percentage of elements with value in an interval [a,b], for any a and b
- **Discrete variable** $Prob(y_i) = \frac{\text{Number of elements}\{e, y(e)=y_i\}}{N} = \frac{N_i}{N}$
- **Continuous variable:** $Prob(a, b) = P(a < y < b) = \int_a^b f(y)dy$
- $f(y)$: Density function

2.2 Some technical definitions

- **Target Population:** population intended to be investigated(Sampled)
- **Sampling Distribution:** Population effectively sampled

- Sampling units: Non-overlapping collections of elements that cover the population effective sampled(SUs)
- **Frame:** List, or any technical device which provides sampling units, or access to sampling units
- **Sample:** Collection of sampling units selected from a frame

3 Design of Sample Survey

3.1 General Procedure:

- Identify target survey group
- Develop questions
- Pilot or test the questions/survey
- Determind the method of conducting the survey
- Conduct the survey
- Use an appropriate analysis technique to analyze the information collected.

3.2 How to select a sample

3.2.1 Cencus

Complete survey of a population.

3.2.2 Probability, or random sampling

Method: Random sampling
Different probability sampling designs have two things in common:

- Every element in the population has a known nonzero probability(not necessarily equal) of being sampled
- involves random selection at same point

E.g Simple random, systematic, Cluster, Double, Stratified

- **Simple Random Samplings(SRS)** sample is selected "completely at random". No special constraints on the sample are imposed. An "unbiased" sample.

- **Stratified Sampling** The population is divided into subpopulations(strata). Random sample is selected from every stratum. Constraint imposed: stratification. i.e the population is divided into different subgroups and pick samples with specified ratio. divide into groups and randomly pick parts from each of the subgroups.**Homogeneity within subgroups**

- **Cluster Sampling(One Stage)** The population is divided into large number of (small) groups(clusters), equal or non-equal. Clusters are selected "at random". Sample: All elements from selected clusters. Constraint imposed: clustering. i.e divided into subgroups and pick whole groups. **Heterogeneity within subgroups**

- **Cluster Sampling(Two-stage)** The population is divided into large number of (bigger) groups(Cluster). **First stage:** sample of clusters selected. **Second:** sample of elements from each selected cluster. Sample: all selected elements. Constraint imposed: clustering. i.e divided into subgroups and pick some samples by ratios from some subgroups.

- **Clustering Sampling(Multi-stage)** The population is divided into clusters on several levels/stages - primary sample units(PSUs), secondary sampling units, ternary sampling units, ...Sampling is performed at every stage. Sample: all sampling units selected at the last stage. Constrained imposed: multi-clustering.

- **Double Sampling(Two-phase sampling)** Sample from sample. First phase: a bigger sample and some basic measurements. Second phase: subsample from previously selected sample and more detailed measurements.

- **Systematic Sampling** Elements are selected from an ordered sampling frame. First element is selected at random, subsequent elements follow a predetermined pattern, usually an interval.

- **Composite design** Most large scale surveys are done using cluster sampling combined with stratification, typically by clustering with strata. i.e divided into subgroups,then each subgroups divided into subsubgroups. Pick sample from each subusbgroups.

3.2.3 Nonrandom/nonprobability sampling

Accidental sampling, quota sampling and purposive/judgemental sampling.

E.g **Quota Sampling:** The criteria for selection of elements are based on quotas - assumptions regarding the population of interest. After the quotas are decided, the choice of actual sampling units to fit into quotas is mostly left to the interviewer.

E.g **Snowball Sampling:** used to recruit more subjects into the sample.

E.g other sampling methods include: convenience, judgment(Purposive)

3.3 Other methods of Data Collection

- **Observational Studies:** The researchers simply observe r measure the participants and do not assign any treatments or conditions. Participants are not asked to do anything differently
- **Experiments:** The researchers manipulate something and measure the effect of the manipulation on some outcome of interest. Often participants are randomly assigned to the various conditions or treatments.
- **Confounding variable:** is a variable that both affect the response variable and also is related to the explanatory variable. The effect of a confounding variable on the response variable cannot be separated from the effect of the explanatory variable.
- **Randomized experiments** helps to control the influence of confounding variables.

3.3.1 Sources of Errors in Surveys

Sampling Errors: Due to random sampling(observing a random sample instead of the whole population), controlled by sample design, sample size and error bound.
Non-sampling Erros: Caused by factors other than those related to sample selection. Are not easily identified or quantified.

- imperfect sampling population
- poorly designed questionnaire
- selection bias, sampling bias
- non-response problem
- response error
- systematic error
- processing, editing entering error

Inadequate Frame - coverage error: The sampling design excludes or under-represents a specific group in the sample, deliberately or not. If the group is different, with respect to survey issues, bias will occur.
Selection bias, interview bias: S Sample members are self-selected volunteers, as in voluntary samples. Individuals with strong opinions about the survey issues or those with substantial knowledge will tend to be over-represented, creating bias.
Interviewer error: Occurs when interviewers incorrectly record information; are not neutral or objective; influence the respondent to answer in a particular way; or assume responses based on appearance or other characteristics.

Response Error(inaccurate error): Caused by respondents intentionally or accidentally providing inaccurate responses.

Non-response bias: failure to obtain a response from some unit because of absence, non-contact, refusal, not-able or some other reason.

- Complete non-response i.e no data had been obtained at all from a select unit
- Partial non-response i.e the answer to some questions have not been provided by a selected unit.

3.3.2 Methods of data collection

- Direct measurements
- Personal interviews
- Telephone interviews
- Mailed questionnaires
- Online internet surveys
- Mobile data collection survey
- Mixed-mode survey

3.4 Sampling design process

1. Define Population
2. Determine sampling frame
3. Determine sampling procedure(probability or non-probability sampling)
4. Determine appropriate sample size
5. Execute sampling design

4 Simple Random Sampling

Simple random sampling of size n is the probability sampling design for which a fixed number of n units are selected from a population of N units such that every possible sample of n units has equal probability of being selected.

SRS are EPSEM samples: Equal probability of Selection of Element Method

4.1 Simple Random Sample without replacement

m units are randomly selected from a population of size N without replacement.
All sample of size n have the same probability of being selected. There are $\binom{N}{n}$ SRS.

4.1.1 Table of random numbers

Table of random numbers: List of digits produced by an RNG Convenient for manual/field/small size problem sampling; repeatable, easy to use.
General use of the table

1. Assign certain digits, or groups of digits to the events A1, A2,... you want to simulate, depending on P(A1), P(A2)
2. Decide how you will read the table, that is, select some digits from the table
3. Read the table, and see which one of the events has occured
4. Read the table from left to right, starting with the first row. Use first 4 digit out of every group of 5 digits until 8 elements are selected.

5 Inference of SRS

N = Population size
Y = Population characteristic
elements in population : $\{u_1, u_2, \cdots, u_N\}$
$\mu = \frac{1}{N}\sum_{i=1}^{N} u_i$
$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(u_i - \mu)^2 = \frac{1}{N}[\sum \mu_i^2 - N\mu^2]$

5.1 Probability

5.1.1 SRS

The probability that sample S is selected:

$$p(S) = \frac{1}{\binom{N}{n}}$$

5.1.2 SRSWR

$$P[y_i = u] = \frac{1}{N}$$

5.2 Inclusion probability of SRS

$$P[y_i = u_i, y_j = u_j] = P[y_u = u_i]P[y_j = u_j|y_i = u_i] = \frac{1}{N} \cdot \frac{1}{N-1}$$

$$\begin{aligned}
E[y_i, y_j] &= \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j P[y_i = u_i, y_j = u_j] \\
&= \sum_{i=1}^{N}\sum_{i\neq j} u_i u_j \frac{1}{N(N-1)} \\
&= \frac{1}{N(N-1)} \sum_{i=1}^{N}\sum_{j\neq i}^{N} u_i u_j \\
\sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j &= \sum_{i=1}^{N}\sum_{j=i}^{N} u_i u_j + \sum_{i=1}^{N}\sum_{j\neq i}^{N} u_i u_j \\
\implies \sum_{i=1}^{N}\sum_{j\neq i}^{N} u_i u_j &= \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j - \sum u_i^2 \\
E[y_i, y_j] &= \frac{1}{N(N-1)}[(\sum u_i)^2 - \sum u_i^2] \\
&= \frac{1}{N(N-1)}[(N\mu)^2 - \sum u_i^2] \\
N\sigma^2 &= \sum u_i^2 - N\mu^2 \implies \sum u_i^2 = N\sigma^2 + N\mu^2 \\
E[y_i, y_j] &= \frac{1}{N(N-1)}[N^2\mu^2 - N\sigma^2 - \sum u_i^2]
\end{aligned}$$

5.2.1 Covariance under SRS

$$\begin{aligned}
cov(y_i, y_j) &= E[(y_i - \mu)(y_j - \mu)] \\
&= E[y_i y_j] - \mu^2 \\
&= \frac{1}{N(N-1)}[N^2\mu^2 - N\sigma^2 - \sum u_i^2] - \mu^2 \\
&= \frac{1}{N(N-1)}[N^2\mu^2 - N\sigma^2 - \sum u_i^2 - \mu^2 \cdot (N^2 - N)] \\
&= \frac{1}{N(N-1)}[N^2\mu^2 - N\sigma^2 - \sum u_i^2 - N^2\mu^2 + N\mu^2] \\
&= \frac{N\sigma^2}{N(N-1)} = \frac{-\sigma^2}{N-1}
\end{aligned}$$

5.3 Inference of Sample mean μ

5.3.1 $\hat{\mu}$ is unbiased:

$$E[\hat{\mu}] = E[\bar{y}] = E[\frac{1}{N}\sum y_n] = \frac{1}{N}\sum_{i=1}^{N} E[y_i] = \frac{1}{n} n\mu$$

5.3.2 Variance of $\hat{\mu}$ under SRSWR

$$\begin{aligned} Var(\hat{\mu}) = Var(\bar{y}) &= Var(\frac{1}{N}\sum y_i) \\ &= \frac{1}{n^2}\sum Var(y_i) = \frac{1}{n^2}\sum Var(y_i) \\ &= \frac{1}{n^2}\sum \sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

5.3.3 Variance of $\hat{\mu}$ under SRS

$$\begin{aligned} Var(\hat{\mu}) = Var(\bar{y}) &= Var(\frac{1}{N}\sum y_i) \\ &= \frac{1}{n^2}[\sum Var(y_i) + \sum_{i=1}\sum_{j\neq i} cov(y_i, y_j)] \\ &= \frac{1}{n^2}[\sum \sigma^2 + \sum\sum_{j\neq i} \frac{-\sigma^2}{N-1}] \\ &= \frac{1}{n^2}[N\sigma^2 - \frac{-\sigma^2}{N-1}\frac{n(n-1)}{2}] = \frac{N-n}{N-1}\frac{\sigma^2}{n} \end{aligned}$$

if we want different i and j, we actually pick two element from the population and have $\binom{n}{2}$ ways of picking such samples.

5.3.4 Summary of population mean $\mu = \bar{y}$

$$E(\bar{y}) = E(\hat{mu}) = \mu$$

$$var(\bar{y}) = \begin{cases} \frac{\sigma^2}{n} & \text{SRSWR} \\ \frac{N-n}{N-1}\frac{\sigma^2}{n} & \text{SRS} \end{cases}$$

$$\widehat{var(\bar{y})} = \begin{cases} \frac{S^2}{n} & \text{Unbiased in SRSWR} \\ (1-\frac{n}{N})\frac{S^2}{n} & \text{unbiased in SRS} \end{cases}$$

5.4 Inference on population variance σ^2

We consider $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$

$$\begin{aligned} E[s^2] &= \frac{1}{n-1}E[\sum(y_i-\bar{y})^2] \\ &= \frac{1}{n-1}\sum[E(y_i-\bar{y})^2] \\ &= \frac{1}{n-1}\sum E([(y_i-\mu)-(\bar{y}-\mu)]^2) \\ &= \frac{1}{n-1}\sum E([(y_i-\mu)^2 - 2(y_i-\mu)(\bar{y}-\mu) + (\bar{y}-\mu)^2] \\ &= \frac{1}{n-1}(\sum E[(y_i-\mu)^2] + \sum E[(\bar{y}-\mu)^2] \\ &= \frac{1}{n-1}[n\sigma^2 - nVar(\bar{y})] \end{aligned}$$

5.4.1 Estimation of σ^2 Under SRSWR

$$Var(\bar{y}) = \frac{\sigma^2}{n}$$
$$E[s^2] = \frac{1}{n-1}[n\sigma^2 - nVar(\bar{y})] = \frac{1}{n-1}[n\sigma^2 - n\sigma^2/n] = \sigma^2$$

5.4.2 Estimation of σ^2 Under SRS

$$Var(\bar{y}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$
$$E[s^2] = \frac{1}{n-1}[n\sigma^2 - nVar(\bar{y})] = \frac{1}{n-1}[n\sigma^2 - n\frac{N-n}{N-1}\frac{\sigma^2}{n}]$$
$$= \frac{1}{n-1}\sigma^2[\frac{Nn-n-N+n}{N-1}] = \frac{1}{n-1}\sigma^2\frac{N(n-1)}{N-1} = \frac{N}{N-1}\sigma^2$$
$$\widehat{Var(\hat{\mu})} = (1-\frac{n}{N})\frac{s^2}{n}$$
$$B_{\hat{\mu}} = 2\sigma_{\hat{\mu}} = 2\sqrt{(1-\frac{n}{N})\frac{s^2}{n}}$$

5.4.3 Summary of Unbiased Estimation of σ^2

$$\hat{\sigma^2} = \begin{cases} S^2 & \text{SRSWR} \\ \frac{N-1}{N}S^2 & \text{SRSS} \end{cases}$$

5.5 Inference of Population Total τ

$$\tau = N\,u \implies \hat{\tau} = N\hat{\mu} = N\bar{y}$$
$$var(\hat{\tau}) = var(N\bar{y}) = N^2 var(\bar{y}) = N^2\frac{N-n}{N-1}\frac{\sigma^2}{n}$$
$$\widehat{var(\hat{\tau})} = \widehat{var(N\bar{y})} = N^2\widehat{var(\bar{y})} = N^2(1-\frac{n}{N})\frac{s^2}{n}$$

5.6 Inference of Population proportion p

For $y_1, y_2, \cdots, y_N, y_i = 1, 0$