STA304/1003 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

Lecture 3 - Part I: Introduction to Probability Sampling

Ramya Thinniyam

May 22, 2014

Indicator Variables

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}$$

In Sampling, we use this RV:

$$Z_i = I(\textit{unit } i \textit{ is in sample}) = \begin{cases} 1, & \text{if unit } i \text{ is in the sample} \\ 0, & \text{otherwise} \end{cases}$$

- $Z_1, \ldots, Z_N$: n RVs will take on the value 1 and remaining $N - n$ will be 0

Properties of Z_i

i in sample, choose (n-1) out of (N-1) in sample

1 ▸ $P(Z_i = 1) = P(\text{ith unit is in the sample}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$

total # ways

2 ▸ $E(Z_i) = \frac{n}{N} = 1P(Z_i=1)+0P(Z_i=0)=P(Z_i=1)$

3 ▸ For $i \neq j$,

$$Z_i = \begin{cases} 1, & \text{wp } P(Z_i=1) \\ 0, & \text{wp } P(Z_i=0)=1-P(Z_i=1) \end{cases}$$

$$P(Z_iZ_j = 1) = P(i \text{ and } j \text{ are in the sample}) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

$P(Z_iZ_j=1)$

4 ▸ For $i \neq j$, $E(Z_iZ_j) = \frac{n(n-1)}{N(N-1)}$

5 ▸ $V(Z_i) = \frac{n(N-n)}{N^2} = Var(Z_i) = E(Z_i^2) - (E(Z_i))^2 = E(Z_i) - (E(Z_i))^2$

$= \frac{n}{N} - \frac{n^2}{N^2} = \frac{n(N-n)}{N^2}$

▸ For $i \neq j$, $Cov(Z_i, Z_j) = \frac{n(n-N)}{N^2(N-1)}$

since $Z_i = Z_i^2$

$E(Z_i) = E(Z_i^2)$ b/c $0^2=0$ & $1^2=1$

Probability Samples

Use Probability Sampling to reduce selection bias and obtain representative samples.

In a Probability Sample, each unit has a known probability of selection.

- Population size = N, Sample size = n
- With a good design, only need small samples to make inferences about large populations
- Use random number table or random number generators (in R) to select units

Types of Probability Samples

1. Simple Random Sample (SRS): Every possible sample of size n has an equal chance of being selected

   • Elements selected randomly
   Ex. balls in urn, numbers in hat - mix the units of the population then randomly select

2. Stratified Random Sample: Population is divided into *strata*, subgroups and a SRS from each stratum is taken independently of other strata.

   • Elements within each strata selected randomly
   • Elements in same strata tend to be similar - increases precision. Strata should be mutually exclusive
   Ex. Political survey - divide by minority groups (race/ethnicity/religions, etc) and sample according to their proportion in the population

3. Cluster Sample: Observations in population put into larger sampling units called *clusters* and take SRS of some clusters and then subsample or sample all members in a cluster. Can have more than one level/cluster - Multi-stage Cluster Sampling.

   • Clusters selected randomly
   • Usually used when you don't have a list of population but can contact them through clusters
   Ex. Absences of Primary School Children - sample schools, then classes, etc

4. Systematic Sample: A starting point is randomly chosen from the population list and then every kth unit is selected to be in the sample.

   • Starting point selected randomly
   • Elements in sample are equally spaced on population list - be careful of patterns hidden in interval
   • Don't have to generate n random numbers - saves time, usually more efficient
   Ex. Want to sample 20 out of 100 customers - sample every 5th customer. $k = \frac{N}{n}$

Setup for Probability Sampling

Assume for now: target population = sampled population, complete sampling frame, no missing data/non-response, no measurement error

- Finite population with N units : $\mathcal{U} = \{1, 2, \ldots, N-1, N\}$
- Sample has n units : $\mathcal{S}$
- Each possible sample, $\mathcal{S}$, has known probability of selection, $P(\mathcal{S})$
- Each unit in the sample has a known probability of selection, $\pi_i = P(\textit{unit i is in sample})$
- π_i known before sampling , $\pi_i > 0$
- Can quantify how often samples will meet certain criteria : doesn't mean each sample is representative

What is random and what is NOT?

y_i $=$ characterisic/variable measured on unit i

- y_i fixed, NOT random
- Selection of units is random
- We sample and then record values
- Randomness is in the selection of unit i that generates y_i

Sampling Distribution: distribution of the statistic - distibution of different values of the statistic obtained by taking all possible samples in the population. (discrete probability distribution).

Estimation

<u>Aim</u>: Estimate Population Total: $t = \sum_{i=1}^{N} y_i$

One possible estimate: $\hat{t} = N\bar{y}_{\mathcal{S}}$

$\bar{y}_{\mathcal{S}} =$ average value of $y's$ in sample $\mathcal{S}$

Sampling distribution can be obtained if we know entire population and sampling distribution calculated as :

$$P(\hat{t} = k) = \sum_{\mathcal{S}:\hat{t}_{\mathcal{S}}=k} P(\mathcal{S})$$

- Expected Value: $E(\hat{t}) = \sum_{\mathcal{S}} \hat{t}_{\mathcal{S}} P(\mathcal{S}) = \sum_k kP(\hat{t} = k)$
- Bias of estimator: Bias$[\hat{t}]$ = $E(\hat{t}) - t$.
  - Estimator is called unbiased if Bias$[\hat{t}]$ = 0 ie. $E(\hat{t}) = t$
  - This bias is not the same as selection/measurement bias
- Variance: $V(\hat{t}) = \sum_{\mathcal{S}} [\hat{t}_{\mathcal{S}} - E(\hat{t}_{\mathcal{S}})]^2 P(\mathcal{S})$
  Called precise if variance is small.
- Mean Squared Error (MSE) : $MSE(\hat{t}) = E[(\hat{t} - t)^2] = V(\hat{t}) + [Bias(\hat{t})]^2$
  Called accurate if MSE is small.

SRS sampling distribution of $\hat{t} = N \cdot \bar{y}_s$

Example: Sample without replacement - $N = 8, n = 2$

Find the sampling distribution of t and its mean and variance.

i	1	2	3	4	5	6	7	8
y_i	1	2	7	7	10	12	19	21

$\sum y_i =$	3	8	9	11	12	13	...
# samples	1	2	2	1	1	1	...

Answer:

Total of $\binom{8}{2} = 28$ possible samples. each sample has prob $\frac{1}{28}$.

k	12	32	36	44	48	52	56	68	76	80	84	88	92	104	112	116	124	132	160
$P(\hat{t} = k)$	1/28	2/28	2/28	1/28	1/28	1/28	2/28	2/28	2/28	1/28	1/28	2/28	1/28	2/28	2/28	1/28	2/28	1/28	1/28

Why 12?

$\hat{t}_1 = N\bar{y}_{s_1} = 8 \cdot \frac{3}{2} = 12$

$$E(\hat{t}) = \sum_k kP(\hat{t} = k) = 79 \quad \text{and} \quad t = \sum_{i=1}^{8} y_i = 79$$

So $\hat{t}$ is an unbiased estimator of t.

$$Var(\hat{t}) = E(\hat{t}^2) - [E(\hat{t})]^2 = 7505.7143 - 79^2 \doteq 1264.7143$$