DATA 606: Statistical Methods in Data Science

—— Simple probability samples

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 2

Prob sampling

Basics:

- Each unit in the population has a known probability of selection.
- Some randomization mechanism is used to choose the specific units to be included in the sample.

Basic sampling frameworks: simple random sample, stratified sample, cluster sample and systematic sample.

Prob sampling

- *Simple random sample (SRS):* an SRS of size n is taken when every possible subset of n units in the population has the same chance of being the sample.
- *Stratified random sample:* the population is divided into subgroups called strata. Then an SRS is selected from each stratum, and the SRS's in the strata are selected independently.
- *Cluster random sample:* observation units in the population are aggregated into larger sampling units, called clusters. Then taking SRS from clusters and subsample the units.
- *Systematic random sample:* a starting point is chosen from a list of population members using a random number. That unit, and every kth unit thereafter, is chosen to be in the sample.

Example 1 (Time spent on grading)

Suppose you want to estimate the average amount of time that professors at your university say they spent grading homework in a specific week.

- SRS: construct a list of all professors and randomly select n of them to be your sample.
- Stratified sample: classify faculty by college: engineering, business, nursing, and fine arts. Then take an SRS of faculty in each college.
- Cluster sample: randomly select 10 out of 30 buildings on campus and survey the professors inside these buildings.
- Systematic sample: selecting an integer at random between 1 and 20; if the random integer is 16, say, then you would include professors in positions 16, 36, 56, and so on, in the list.

Framework

- Finite population of N units:

$$U = \{1, 2, \ldots, n\}.$$

A sample S contains a subset of U.

Example 2

Let $U = \{1, 2, 3, 4\}$, then a sample containing two elements of U could be

$$S_1 = \{1, 2\}, \quad S_2 = \{1, 3\}, \quad S_3 = \{1, 4\}$$
$$S_4 = \{2, 3\}, \quad S_5 = \{2, 4\}, \quad S_6 = \{3, 4\}.$$

In a SRS framework, we know $\mathbf{P}(S_1) = \cdots = \mathbf{P}(S_6) = \frac{1}{6}$. However, for some survey, it could be $\mathbf{P}(S_1) = \frac{1}{2}$, $\mathbf{P}(S_2) = \frac{1}{6}$, $\mathbf{P}(S_3) = \frac{1}{3}$ and $\mathbf{P}(S_4) = \mathbf{P}(S_5) = \mathbf{P}(S_6) = 0$.

Framework

- Probability of inclusion

$$\pi_i = \mathbf{P}(\text{unit } i \text{ is in the sample}).$$

Example 3

In the example 2, for SRS $\pi_1 = \mathbf{P}(S_1) + \mathbf{P}(S_2) + \mathbf{P}(S_3) = \frac{1}{2}$. However, for the second designed survey, $\pi_1 = \mathbf{P}(S_1) + \mathbf{P}(S_2) + \mathbf{P}(S_3) = \frac{2}{3}$.

Framework

Example 2.2 from book

To illustrate these concepts, let's look at an artificial situation in which we know the value of y_i for each of the $N = 8$ units in the whole population. The index set for the population is

$$\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

The values of y_i are

i	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

There are 70 possible samples of size 4 that may be drawn without replacement from this population; the samples are listed in file samples.dat on the website. If the sample consisting of units $\{1, 2, 3, 4\}$ were chosen, the corresponding values of y_i would be 1, 2, 4, and 4. The values of y_i for the sample $\{2, 3, 6, 7\}$ are 2, 4, 7, and 7. Define $P(\mathcal{S}) = 1/70$ for each distinct subset of size four from $\mathcal{U}$. As you will see after you read Section 2.3, this design is an SRS without replacement. Each unit is in exactly 35 of the possible samples, so $\pi_i = 1/2$ for $i = 1, 2, \ldots, 8$.

Framework

Usually, after sampling, we are interested in some characteristic (y_i) of unit i (could be income, age, marriage status, etc.). Through the sample, we want to estimate

- Population total: $t_U = \sum_{i=1}^{N} y_i$.
- Population average (mean): $\bar{y}_U = \frac{t_U}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$.
- Population variance: $V = \frac{\sum_{i=1}^{N} (y_i - \bar{y}_U)^2}{N-1}$.

When we have a sample S of size n, we would have

$$\bar{y}_S = \frac{\sum_{i \in S} y_i}{n}, \quad \hat{t}_S = N \cdot \bar{y}_S, \quad v = \frac{\sum_{i \in S} (y_i - \bar{y}_S)^2}{n-1}.$$

Framework

For different samples, we usually have different $\hat{t}_S$ (estimator). As the sample is selected randomly, $\hat{t}_S$ is random. We call the distribution of $\hat{t}_S$ (or maybe other statistics) the *sampling distribution*:

$$\mathbf{P}(\hat{t}_S = k) = \sum_{S:\hat{t}_S=k} \mathbf{P}(S)$$

The expected value of $\hat{t}_S$ is

$$\mathbf{E}[\hat{t}_S] = \sum_S \hat{t}_S \mathbf{P}(S) = \sum_k k\mathbf{P}(\hat{t}_S = k).$$

Framework

Example 2.3 from book

EXAMPLE 2.3 The sampling distribution of $\hat{t}$ for the population and sampling design in Example 2.2 derives entirely from the probabilities of selection for the various samples. Four samples ($\{3,4,5,6\}$, $\{3,4,5,7\}$, $\{3,4,6,7\}$, and $\{1,5,6,7\}$) result in the estimate $\hat{t} = 44$, so $P\{\hat{t} = 44\} = 4/70$. For this example, we can write out the sampling distribution of $\hat{t}$ because we know the values for the entire population.

k	22	28	30	32	34	36	38	40	42	44	46	48	50	52	58
$P\{\hat{t} = k\}$	$\frac{1}{70}$	$\frac{6}{70}$	$\frac{2}{70}$	$\frac{3}{70}$	$\frac{7}{70}$	$\frac{4}{70}$	$\frac{6}{70}$	$\frac{12}{70}$	$\frac{6}{70}$	$\frac{4}{70}$	$\frac{7}{70}$	$\frac{3}{70}$	$\frac{2}{70}$	$\frac{6}{70}$	$\frac{1}{70}$

Framework

- The *estimation bias* of $\hat{t}_S$ is

$$\mathrm{Bias}(\hat{t}_S) = \mathbf{E}[\hat{t}_S] - t_U.$$

If $\mathrm{Bias}(\hat{t}_S)=0$, then we say $\hat{t}_S$ is unbiased.

- The *sample variance* of $\hat{t}_S$ is

$$\mathrm{Var}(\hat{t}_S) = \mathbf{E}[(\hat{t}_S - \mathbf{E}[\hat{t}_S])^2].$$

If $\mathrm{Var}(\hat{t}_S)$ is very small, then we say $\hat{t}_S$ is precise.

- The *mean square error* (MSE) of $\hat{t}_S$ is

$$\mathrm{MSE}(\hat{t}_S) = \mathbf{E}[(\hat{t}_S - t_U)^2].$$

If $\mathrm{MSE}(\hat{t}_S)$ is very small, we say $\hat{t}_S$ is accurate.

Simple random sample (SRS)

- Two ways: with replacement and without replacement.
- To take a sample of size n from a population of size N, there are in total $\binom{N}{n}$ samples could be possibly selected.
- Each sample could be picked with probability

$$\frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}.$$

- The probability that unit i is included in the sample is

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

SRS

How to sample SRS?

- Need a sampling frame: a list of all the units in the population.
- Number these units.
- Use computer to generate "random" numbers (uniform distribution on $[0, 1]$).
- Select the n smallest numbers.

SRS

How to sample SRS?

- Need a sampling frame: a list of all the units in the population.
- Number these units.
- Use computer to generate "random" numbers (uniform distribution on $[0, 1]$).
- Select the n smallest numbers.

Example: select 4 out of 10.

unit i	1	2	3	4	5	6	7	8	9	10
random number	0.837	0.636	0.465	0.609	0.154	0.766	0.821	0.713	0.987	0.469

The smallest 4 random numbers are: 0.154, 0.465, 0.469 and 0.609, the corresponding units are $\{3, 4, 5, 10\}$.

- To estimate the population mean $\bar{y}_U$, we use the sample mean $\bar{y}_S$.
- Note: for different samples, we have different $\bar{y}_S$. As such

$$\mathrm{Var}(\bar{y}_S) = \frac{V}{n}\left(1 - \frac{n}{N}\right).$$

- The population variance V is usually unknown, it is estimated using the sample variance

$$v = \frac{1}{n-1}\sum_{i \in S}(y_i - \bar{y}_S)^2.$$

- To sum up, an estimate for $\mathrm{Var}(\bar{y}_S)$ is

$$\hat{\mathrm{Var}}(\bar{y}_S) = \frac{v}{n}\left(1 - \frac{n}{N}\right).$$

- The population total is $t = \sum_{i=1}^{N} y_i = N \cdot \bar{y}_U$.
- Its estimate is given by $\hat{t} = N \cdot \bar{y}_S$.
- From the previous slide, we know

$$\mathrm{Var}(\hat{t}) = \frac{V}{n}\left(1 - \frac{n}{N}\right) N^2.$$

- An estimate of $\mathrm{Var}(\hat{t})$ is

$$\hat{\mathrm{Var}}(\hat{t}) = \frac{v}{n}\left(1 - \frac{n}{N}\right) N^2.$$

95% confidence interval: understand it correctly!

**If we take samples again and again and construct the interval as per our procedure, 95% of the resulting intervals could cover the true value.

95% confidence interval: understand it correctly!
**If we take samples again and again and construct the interval as per our procedure, 95% of the resulting intervals could cover the true value.

*If we are able to generate all the possible samples, we can calculate the exact confidence interval.

- As per the central limit theorem ,

$$\frac{\bar{y}_S - \bar{y}_U}{\sqrt{\left(1 - \frac{n}{N}\right) \frac{V}{n}}} \sim N(0, 1).$$

- When replacing V with the its estimate v,

$$\frac{\bar{y}_S - \bar{y}_U}{\sqrt{\left(1 - \frac{n}{N}\right) \frac{v}{n}}} \sim t_{n-1}.$$

- The resulting $1 - \alpha\%$ confidence interval of $\bar{y}_U$

$$\left[\bar{y}_S - t_{\alpha/2, 2n-1}\sqrt{\left(1 - \frac{n}{N}\right) \frac{v}{n}},\ \bar{y}_S + t_{\alpha/2, 2n-1}\sqrt{\left(1 - \frac{n}{N}\right) \frac{v}{n}}\right].$$

$t_{\alpha/2, 2n-1}$: $(1 - \alpha/2)\%$ percentile of a t distribution with DOF $2n - 1$.

- **Specify the tolerable error**

$$\mathbf{P}(|\bar{y}_S - \bar{y}_U| < e) = 1 - \alpha,$$

where e is called the margin of error.

- **Find an equation**

$$\mathbf{P}\left(\frac{|\bar{y}_S - \bar{y}_U|}{\sqrt{\left(1 - \frac{n}{N}\right)\frac{V}{n}}} < \frac{e}{\sqrt{\left(1 - \frac{n}{N}\right)\frac{V}{n}}}\right) = 1 - \alpha,$$

$$\longrightarrow \quad \frac{e}{\sqrt{\left(1 - \frac{n}{N}\right)\frac{V}{n}}} = z_{\alpha/2},$$

$$\longrightarrow \quad n = \frac{z_{\alpha/2}^2 V}{e^2 + \frac{z_{\alpha/2}^2 V}{N}}.$$

Some methods for estimating V (before you can conduct the survey):

- Use sample quantities obtained when pretesting your survey.
- Use previous studies or data available in the literature.
- If nothing else is available, guess the variance.

$w_i = \frac{N}{n}$

Discussed but not covered in slides
- Sampling Weight