Stats 100B Chapter 7 Survey Sampling

Chapter 7.1 Introduction

Survey sampling is a statistical tool used to obtain information about a large population by examining only a small fraction of the population.

- Applications in many fields.
  - Government surveys such as health surveys/household incomes.
  - Estimate number of fish in a large lake.
  - Estimate number of homeless people in LA county.
- We discuss sampling techniques that are probabilistic in nature.
- Each member of the population has a specified prob. of being included in the sample.
- The actual composition of the sample is random.

Chapter 7.2 Population Parameters

A population of size N, each member or unit has a numerical value (x-value), say, $x_1, \ldots, x_N$.

- popu. mean: $\mu = \frac{1}{N}(x_1 + \ldots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$
- popu. variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$
- popu. standard deviation (sd): $\sigma = \sqrt{\sigma^2}$

Chapter 7.3 Simple Random Sampling (SRS)

SRS: Each sample of size n has the same probability of occurrence.

- There are total $\binom{N}{n}$ possible sample of size n taken **without replacement**.
- So each sample has prob= $1/\binom{N}{n}$ to be selected.

7.3.1 Expectation and Variance of the sample mean

For an SRS of size n ($< N$), let $X_1, \ldots, X_n$ be the values of the sampled units.

- Each x_i is a fixed number, but each X_i is a r.v.
- x_i is the value of the ith unit of the population, which is fixed.
- X_i is the value of the ith unit of the sample, which is random.
- $X_1, \ldots, X_n$ have the same distribution, but they are NOT independent. (Why?)

- What are $E(X_i)$ and $V(X_i)$?
- Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$

We don't know popu. mean μ and so use sample mean $\bar{X}$ to estimate μ.

- Sample mean $\bar{X}$ is a r.v., which has a probability distribution.
- Its probability distribution is called its **sampling distribution**.

Example A. A population of $N = 393$ hospitals. $x_i = \#$ patients discharged for the ith hospital in January 1968; see `hospitals.txt` and Figure 7.1.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = 814.6, \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} = 588.97$$

(textbook: $\sigma = 589.7$ is a mistake caused by dividing $N-1$ instead of N)

What is the sampling distribution of $\bar{X}$ of an SRS of size n?
If $n = 1$, there are 393 SRS. If $n = 16$, there are $\binom{393}{16} \approx 1.14 \times 10^{28}$ SRS of size 16.

Use simulation

- draw many samples of size n
- compute the sample mean for each sample
- form a histogram of the collection of sample means

Figure 7.2. Histograms of sample means of 500 SRS of size (a) n=8, (b) n=16, (c) n=32, (d) n=64

- All the histograms are centered about the popu mean $\mu = 814.6$.
- As n increases, the histograms become less spread out.
- The histograms are nearly symmetric about the mean. (compare with Fig. 7.1 histogram of popu values is not symmetric.)

Theorem A. For SRS,

$$E(\bar{X}) = \mu$$

- Sample mean $\bar{X}$ is an **unbiased estimator** of popu. mean μ.
- When averaging over all possible $\binom{N}{n}$ SRS, the average of $\bar{X}$ is μ,

Q: What is $V(\bar{X})$? Is it $V(\bar{X}) = \sigma^2/n$?

- Recall, if $X_1, \ldots, X_n$ are independent, $V(\bar{X}) =$
- This happens when the sampling were done **with replacement**.
- For SRS without replacement, $X_1, \ldots, X_n$ are NOT independent.

Lemma B. For SRS (without replacement), for $i \neq j$,

$$cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

Theorem B. For SRS (without replacement),

$$V(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

- $\frac{N-n}{N-1} = 1 - \frac{n-1}{N-1}$ is called the **finite popu. correction factor**.
  - n/N is the sampling fraction.
  - If n/N is very small ($n << N$), $1 - \frac{n-1}{N-1} \approx 1$ and $V(\bar{X}) \approx \frac{\sigma^2}{n}$.
  - If $n/N = 1$, $V(\bar{X}) = 0$ (no sampling error).

Standard error of $\bar{X}$, denoted by $\sigma_{\bar{X}}$, is the standard deviation of $\bar{X}$, i.e.,

$$\sigma_{\bar{X}} = \sqrt{V(\bar{X})}$$

7.3.2 Estimation of Population Variance σ^2

Popu variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ vs sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Theorem A. For SRS (without replacement),

$$E(S^2) = \frac{N}{N-1}\sigma^2$$

An **unbiased estimator** of popu variance σ^2 is $(1 - \frac{1}{N})S^2$ as

$$E[(1 - \frac{1}{N})S^2] = \sigma^2$$

7.3.3 Normal Approximation to Sampling Distribution of $\bar{X}$

Recall Chapter 5, CLT: If $X_1, \ldots, X_n$ are i.i.d. r.v.'s with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$,

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \text{ for large } n$$

$$P(\frac{\bar{X} - \mu}{\sqrt{V(\bar{X})}} \leq z) \to \Phi(z), \text{ as } n \to \infty$$

For SRS, N is fixed and $X_1, \ldots, X_n$ are not independent.

$$E(\bar{X}) = \mu, \quad cov(X_i, X_j) = -\frac{\sigma^2}{N-1}, \quad V(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

If n is large and $N >> n$, CLT still holds for SRS.

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0, 1) \text{ for large } n$$

$$\bar{X} \sim N(\mu, \sigma^2_{\bar{X}}) \text{ for large } n$$

where $\sigma_{\bar{X}} = se(\bar{X}) = \sqrt{V(\bar{X})}$ is the standard error of $\bar{X}$.

CLT Simulation (see `hospitals.R, hospitals-out.pdf`)

$$\begin{aligned}
P(|\bar{X}-\mu| \le \delta) &= P(-\delta \le \bar{X}-\mu \le \delta) = P(-\frac{\delta}{\sigma_{\bar{X}}} \le \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \le \frac{\delta}{\sigma_{\bar{X}}}) \\
&\approx P(-\frac{\delta}{\sigma_{\bar{X}}} \le Z \le \frac{\delta}{\sigma_{\bar{X}}}) = \Phi(\frac{\delta}{\sigma_{\bar{X}}}) - \Phi(-\frac{\delta}{\sigma_{\bar{X}}}) = 2\Phi(\frac{\delta}{\sigma_{\bar{X}}}) - 1 \\
P(|\bar{X}-\mu| > \delta) &= 1 - P(|\bar{X}-\mu| \le \delta) \approx 2 - 2\Phi(\frac{\delta}{\sigma_{\bar{X}}})
\end{aligned}$$

Example A. $N = 393$, $\mu = 814.6$, $\sigma = 588.97$. For SRS of size $n = 64$,

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)} = 67.45$$

$$P(|\bar{X}-\mu| > 100) \approx 2 - 2\Phi(100/67.45) = 2 - 2\Phi(1.48) = 2(1 - .9306) = .1389 \approx .14$$

Simulation: Draw $K = 1000$ SRS of size $n = 64$.

- Expect: About 14% of $\bar{X}$'s differed by more than 100 from $\mu = 814.6$.
- We got 0.148 from $K = 1000$ SRS.

Confidence Interval (CI) for popu mean μ

$$\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \approx Z \sim N(0,1) \text{ for large } n$$

$$P(\bar{X} - z(\alpha/2)\sigma_{\bar{X}} \le \mu \le \bar{X} + z(\alpha/2)\sigma_{\bar{X}}) \approx 1 - \alpha$$

where $z(\alpha) = z_p$ with $p = 1 - \alpha$; see Figure 7.3. So $P(-z(\alpha/2) \le Z \le z(\alpha/2)) = 1 - \alpha$.

A $100(1-\alpha)\%$ approximate **confidence interval (CI)** for μ is

$$\bar{X} \pm z(\alpha/2)\sigma_{\bar{X}} = (\bar{X} - z(\alpha/2)\sigma_{\bar{X}}, \bar{X} + z(\alpha/2)\sigma_{\bar{X}})$$

- This CI is random as $\bar{X}$ is random.
- In practice, $\sigma_{\bar{X}}$ is unknown. We estimate it by $S_{\bar{X}} = \sqrt{\frac{S^2}{n}(1 - \frac{n}{N})}$ for SRS.
- For large n (≥ 25 or 30), the difference between $\sigma_{\bar{X}}$ and $S_{\bar{X}}$ is small and can be ignored.

Simulation (Figure 7.4): Construct 95% CI bases $K = 20$ SRS of size $n = 25$. Here $\alpha = .05$, $z(\alpha/2) = z(.025) = 1.96$.

- 95% CI: $\bar{X} \pm 1.96\sigma_{\bar{X}}$ or $\bar{X} \pm 1.96 S_{\bar{X}}$
- margin of error $= 1.96\sigma_{\bar{X}}$ or $1.96 S_{\bar{X}}$.
- Expect: Approx 95% of our CI's contain the true μ.

Discussion: A particular 95% CI for μ is (553.7, 1001.1). What is $P(553.7 \le \mu \le 1001.1)$?