Chapter 2: Survey Sampling

Mathematical Statistics

UIC-DMS

March 12, 2024

Table of Contents

Section 1

Introduction

Survey Sampling

Sample surveys are used to obtain information about a large population. The purpose of **survey sampling** is to reduce the cost and the amount of work that it would take to survey the entire population.

By a small sample
we may judge of the whole piece

Miguel de Cervantes
"Don Quixote"

Familiar Examples of Survey Sampling:

- the cook in the kitchen taking a spoonful of soup to determine its taste
- the brewer needing only a sip of beer to test its quality

History of Survey Sampling

The first known attempt to make statements about a population using only information about part of it was made by the English merchant John Graunt. In his famous tract (Graunt, 1662) he describes a method to estimate the population of London based on partial information. John Graunt has frequently been merited as the founder of demography.

The second time a survey-like method was applied was more than a century later. Pierre Simon Laplace realized that it was important to have some indication of the accuracy of the estimate of the French population (Laplace, 1812).

Survey Sampling: Population Parameters

Suppose that the target population is of size N (N is very large) and a numerical value of interest x_i is associated with i^{th} member of the population, $i = 1, \ldots, N$.

Examples:

- x_i = age, weight, etc.
- $x_i = 1$ if some characteristic is present, and $x_i = 0$ otherwise.

There are two "standard" parameters of population that we are typically interested:

Definition 2.1.1

- Population mean
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$
- Population variance
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

Simple Random Sampling

Note that μ and σ^2 are not random. They are some fixed unknown parameters. We want to estimate them by picking n out of N members of the population and constructing estimates of μ and σ^2 based only on these n members.

The most elementary form of sampling from a population is simple random sampling.

Definition 2.1.2 (Simple Random Sampling)

In Simple Random Sampling, each member is chosen entirely by chance and, therefore, each member has an equal chance of being included in the sample; each particular sample of size n has the same probability of occurrence.

Important Remark:

- We assume that sampling is done without replacement so that each member of the population will appear in the sample at most once.
- There are in total $\binom{N}{n}$ ways to get a sample of size n.
- In modern applications, N is generally very large. Also, n is big but generally much smaller than N.

Simple Random Sampling cont.

Let $X_1, \ldots, X_n$ be the sample drawn from the population.

Important Remark: Each X_i is a random variable:

- X_i is the value of the i^{th} element of the sample that was randomly chosen from the population.
- Since the sampling is done without replacement, the random variables $X_1, X_2, \ldots, X_n$ are not independent.
- x_i is the observed value of the i^{th} member of the population.

Estimate

We will consider the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as an estimate of the population mean μ. Since X_i are random, $\overline{X}_n$ is also random. Distribution of $\overline{X}_n$ is called its sampling distribution. The sampling distribution of $\overline{X}_n$ determines how accurately $\overline{X}_n$ estimates μ: the more tightly the sampling distribution is centered on μ, the better the estimate.

- Our goal: is to investigate the sampling distribution of $\overline{X}_n$

Since $\overline{X}_n$ depends on X_i, let us start with examining the distribution of a single sample element X_i.

Basic Lemma

Lemma 2.1.3

Denote the distinct values assumed by the population members by $\xi_1, \ldots, \xi_m$, $m \leq N$, and denote the number of population members that have the value ξ_i by n_i. Then X_i is a discrete random variable with probability mass function

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N}$$

Also

$$\mathbb{E}[X_i] = \mu \quad \text{Var}[X_i] = \sigma^2$$

$\overline{X}_n$ is an unbiased estimator of μ

Theorem 2.1.4

With simple random sampling,

$$\mathbb{E}\left[\overline{X}_n\right] = \mu$$

This result can be interpreted as follows: "on average" $\overline{X}_n = \mu$

Definition 2.1.5

Suppose we want to estimate a parameter θ by a function $\hat{\theta}$ of the sample $X_1, \ldots, X_n$

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

The estimator $\hat{\theta}$ is called unbiased if $\mathbb{E}[\hat{\theta}] = \theta$

Thus, $\overline{X}_n$ is an unbiased estimator of μ

Summary

- Sample surveys are used to obtain information about a large population
- Population parameters: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- We use sample mean $\overline{X}_n$ to estimate the population mean μ.
- μ is unknown fixed parameter
- $\overline{X}_n$ is random
- Properties of the sample element X_i :

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N} \quad \mathbb{E}[X_i] = \mu \quad \text{Var}[X_i] = \sigma^2$$

- $\overline{X}_n$ is an unbiased estimator of μ

$$\mathbb{E}\left[\overline{X}_n\right] = \mu$$

- Our next goal is to study the sampling distribution of $\overline{X}_n$.

Section 2

Accuracy of estimation of the population mean

In previous lectures, we discussed the basic mathematical framework of survey sampling:

- We have the target population of size N (N is very large).
- A numerical value of interest x_i (age, weight, income, etc) is associated with i^{th} member of the population.
- We are interested in population parameters:
  - Population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- We estimate μ by the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ is a sample drawn from the population using the simple random sampling.

We proved that $\overline{X}_n$ is an unbiased estimate of μ:

$$\mathbb{E}\left[\overline{X}_n\right] = \mu$$

In other words, on average $\overline{X}_n \approx \mu$. Our next goal is to investigate how variable $\overline{X}_n$ is

As a measure of the dispersion of $\overline{X}_n$ about μ, we will use the standard deviation of $\overline{X}_n$, $\sigma_{\overline{X}_n} = \sqrt{\text{Var}\left[\overline{X}_n\right]}$.

Thus, we want to find

$$\text{Var}\left[\overline{X}_n\right] = ?$$

$$\text{Var}\left[\overline{X}_n\right] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} X_i\right]$$

Remark: If sampling were done with replacement then X_i would be independent, and we would have:

$$\text{Var}\left[\overline{X}_n\right] = \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

In simple random sampling, we do sampling without replacement. This induces dependence among X_i. And therefore

$$\text{Var}\left[\overline{X}_n\right] = \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} X_i\right] \neq \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[X_i\right]$$

Recall:

$$\text{Var}\left[\sum_{i=1}^{n}\alpha_i X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\,\text{Cov}\left(X_i, X_j\right)$$

Thus, we have:

$$\text{Var}\left[\overline{X}_n\right] = \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}\left(X_i, X_j\right)$$

So, we need to find $\text{Cov}\left(X_i, X_j\right)$.

Lemma 2.2.6

If $i \neq j$, then the covariance between X_i and X_j is

$$\text{Cov}\left(X_i, X_j\right) = -\frac{\sigma^2}{N-1}$$

Theorem 2.2.7

The variance of $\overline{X}_n$ is given by

$$\text{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Important observations:

- If $n << N$, then

$$\text{Var}\left[\overline{X}_n\right] \approx \frac{\sigma^2}{n} \quad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}$$

  $\left(1 - \frac{n-1}{N-1}\right)$ is called finite population correction.

- To double the accuracy of $\mu \approx \overline{X}_n$, the sample size must be quadrupled
- If σ is small (the population values are not very dispersed), then a small sample will be fairly accurate. But if σ is large, then a larger sample will be required to obtain the same accuracy.

Summary

- The main result of this section is the expression for the variance of $\overline{X}_n$:

$$\text{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

- The corresponding standard deviation

$$\sigma_{\overline{X}_n} = \sqrt{\text{Var}\left[\overline{X}_n\right]}$$

  measures the dispersion of $\overline{X}_n$ about μ.

Section 3

Estimation of the Population Variance

Agenda

- Why do we need to estimate σ?
- How can we estimate σ?
- Summary

The Need of Estimation of σ

We know that the sample mean $\overline{X}_n$ is an unbiased estimate of the population mean μ :

$$\mathbb{E}\left[\overline{X}_n\right] = \mu$$

Moreover, the accuracy of the approximation $\overline{X}_n \approx \mu$ can be measured by the standard deviation of $\overline{X}_n$ (also called "standard error"):

$$\sigma_{\overline{X}_n} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}, \quad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}, \quad \text{if } n \ll N \tag{1}$$

where σ is the population variance

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Q: What is the main drawback of (1)?
A: We can't use (1) since σ is unknown. To use (1), σ must be estimated from the sample $X_1, \ldots, X_n$.

Estimation of σ

It seems natural to use the following estimate

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

However, this estimate is biased.

Theorem 2.3.8

The expected value of $\hat{\sigma}_n^2$ is given by

$$\mathbb{E}\left[\hat{\sigma}_n^2\right] = \sigma^2\frac{Nn - N}{Nn - n}$$

Important Remark:

- Since $\frac{Nn-N}{Nn-n} < 1$, we have $\mathbb{E}\left[\hat{\sigma}_n^2\right] < \sigma^2$.
  Therefore, $\hat{\sigma}_n^2$ tends to underestimate σ^2

Corollary 2.3.9

Since $\mathbb{E}\left[\hat{\sigma}_n^2\right] = \sigma^2\frac{Nn-N}{Nn-n}$,

$$\hat{\sigma}^2_{n,\ unbiased} = \frac{Nn - n}{Nn - N}\hat{\sigma}_n^2$$

is an unbiased estimate of σ^2

Recall that

$$\text{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

In practice, σ is unknown, so we need to estimate it.

Corollary 2.3.10

An unbiased estimate of $\text{Var}\left[\overline{X}_n\right]$ *is*

$$s^2_{\overline{X}_n} = \frac{\hat{\sigma}_n^2}{n}\frac{Nn - n}{Nn - N}\left(1 - \frac{n-1}{N-1}\right)$$

Summary

Things that we've learnt about the estimation of population parameters:

- Population mean μ
  - Unbiased estimate:
    $$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i$$
  - Variance of estimate
    $$\text{Var}\left[\overline{X}_n\right] \equiv \sigma^2_{\overline{X}_n} = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$
  - Estimated variance
    $$\sigma^2_{\overline{X}_n} \approx s^2_{\overline{X}_n} = \frac{\hat{\sigma}_n^2}{n}\frac{Nn - n}{Nn - N}\left(1 - \frac{n-1}{N-1}\right)$$
- Population variance σ
  - Unbiased estimate:
    $$\hat{\sigma}^2_{n,\ \text{unbiased}} = \frac{Nn - n}{Nn - N}\hat{\sigma}_n^2, \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

Conclusion

In simple random sampling, we can not only form estimate of unknown population parameter (e.g. μ), but also obtain the likely size of errors of these estimates. In other words, we can obtain the estimate of a parameter as well as the estimate of the error of that estimate

Section 4

The Normal Approximation to the Sampling Distribution of $\overline{X}$

Agenda

- Normal Approximation (theoretical result)
- Approximation of the Error Probabilities (application 1)
- Confidence Intervals (application 2)
- Example: Hospitals
- Summary

In previous lectures, we found the mean and the variance of the sample mean:

$$\mathbb{E}\left[\overline{X}_n\right] = \mu \quad \text{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Ideally, we would like to know the entire distribution of $\overline{X}_n$ (sampling distribution) since it would tell us everything about the random variable $\overline{X}_n$

Reminder: If $X_1, \ldots, X_n$ are i.i.d. with the common mean μ and variance σ^2, then the sample mean $\overline{X}_n$ has the following properties:

1. $\mathbb{E}\left[\overline{X}_n\right] = \mu, \quad \text{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$
2. CLT:
$$\mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z), \quad \text{as } n \to \infty$$
where $\Phi(z)$ is the CDF of $N(0,1)$

Q: Can we use these results to obtain the distribution of $\overline{X}_n$?
A: No. In simple random sampling, X_i are not independent. Moreover, it makes no sense to have n tend to infinity while N is fixed.

Nevertheless, it can be shown that if n is large, but still small relative to N, then $\overline{X}_n$ is **approximately normally distributed**

$$\boxed{\overline{X}_n \dot{\sim} N\left(\mu, \sigma^2_{\overline{X}_n}\right)} \quad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}}$$

How can we use this results?
Suppose we want to find the probability that the error made in estimating μ by $\overline{X}_n$ is less than $\varepsilon > 0$. In symbols, we want to find

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \leq \varepsilon\right) = ?$$

Theorem 2.4.11

From $\overline{X}_n \dot{\sim} N\left(\mu, \sigma^2_{\overline{X}_n}\right)$ *it follows that*

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \leq \varepsilon\right) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1$$

Confidence Intervals

Let $\alpha \in [0, 1]$

Definition 2.4.12

A $100(1-\alpha)\%$ confidence interval for a population parameter θ is a random interval calculated from the sample, which contains θ with probability $1-\alpha$.

Interpretation:
If we were to take many random samples and construct a confidence interval from each sample, then about $100(1-\alpha)\%$ of these intervals would contain θ.

Our goal: to construct a confidence interval for μ
Let z_α be that number such that the area under the standard normal density function to the right of z_α is α. In symbols, z_α is such that

$$\Phi(z_\alpha) = 1 - \alpha$$

Useful property:

$$z_{1-\alpha} = -z_\alpha$$

Confidence interval for μ

Theorem 2.4.13

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

$$\left(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}\right)$$

That is the probability that μ *lies in that interval is approximately* $1-\alpha$

$$\mathbb{P}\left(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n} \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}\right) \approx 1-\alpha$$

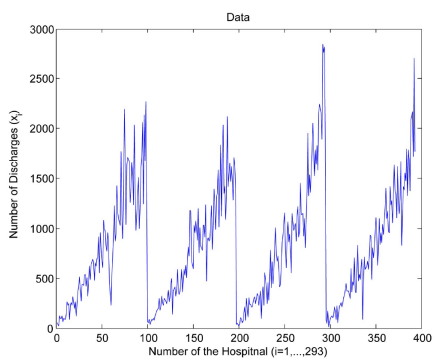
Remarks:

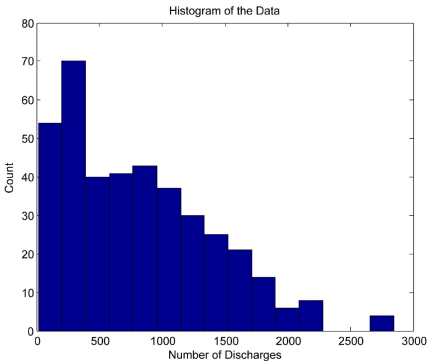
- This confidence interval is random. The probability that it covers μ is $(1-\alpha)$
- In practice, $\alpha = 0.1, 0.05, 0.01$ (depends on a particular application)
- Since $\sigma_{\overline{X}_n}$ is not known (it depends on σ), $s_{\overline{X}_n}$ is used instead of $\sigma_{\overline{X}_n}$

Example: Hospitals

Data: Herkson (1976):

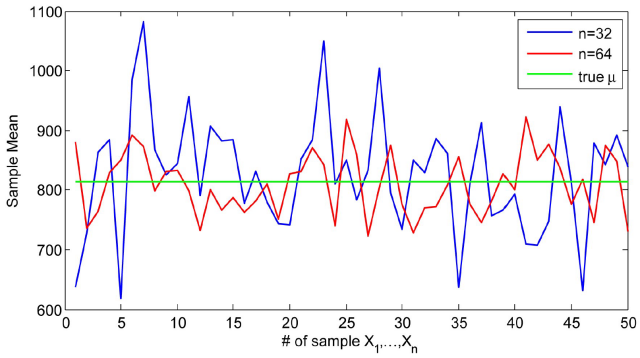
- The population consists of $N = 393$ short-stay hospitals
- Let x_i be the number of patients discharged from the i^{th} hospital during January 1968.





Example: Hospitals

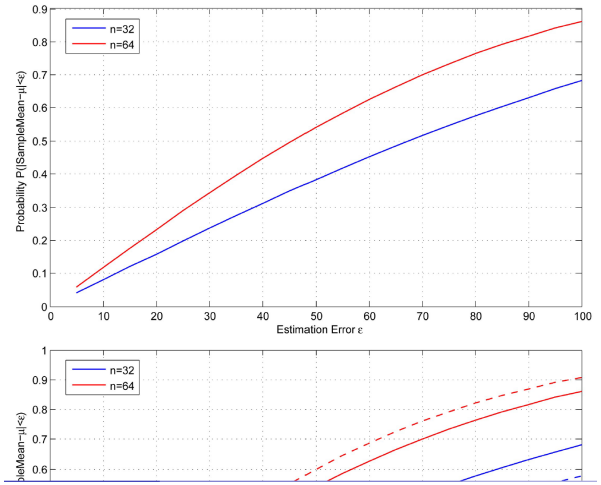
- Population mean $\mu = 814.6$, and population variance $\sigma^2 = (589.7)^2$
- Let us consider two case $n_1 = 32$ and $n_2 = 64$.



- True std of $\overline{X}_n$: $\sigma_{\overline{X}_n} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$, $\sigma_{\overline{X}_{32}} = 100$, $\sigma_{\overline{X}_{64}} = 67.5$

Example: Hospitals

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \leq \varepsilon\right) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1$$

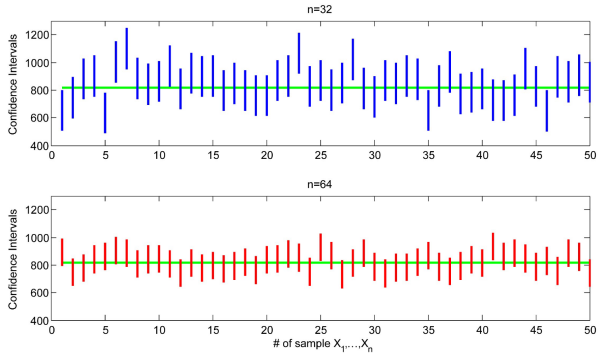


Example: Hospitals

$100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}\right)$$

$\alpha = 0.1$:



Interval width: 329.1 for $n = 32$ and 222.2 for $n = 64$

Summary

- The sample mean is approximately normal

$$\boxed{\overline{X}_n \dot{\sim} N\left(\mu, \sigma^2_{\overline{X}_n}\right)} \quad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$

- Probability of error

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \leq \varepsilon\right) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1$$

- $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}\right)$$