Chapter 7: Survey Sampling (Rice 7.1–7.3)

Donghyun Ko

May 27, 2026

Survey sampling uses a *randomly selected sample* to infer population-level quantities (mean, total, proportion, variance) while correctly accounting for *sampling randomness*. In Rice 7.1–7.3, the main message is that *sampling design* (with vs. without replacement) directly determines the bias/variance of estimators and their *standard errors*.

Contents

1 From "measure everyone" to "sample and infer"

In order to learn how a characteristic varies in a population (e.g., heights of all 20-year-olds in the U.S.), we could measure the entire population. In practice, this is usually impossible due to time/budget constraints. Instead, we collect a representative (random) **sample** and use it to estimate population features. Representative sample is a subset of subjects/objects from the population, which is selected at random!

Population is the collection of subjects/objects on which we measure the characteristic of interest. It is defined by the research question. We can use the distribution of the characteristic in the sample to make inferences about how the characteristic varies in the population!

Example

Identify the population:

- "Distribution of height of 20-year-olds in the U.S." ⇒ all 20-year-olds in the U.S.
- "Acreage of wheat in a state" ⇒ all land units/farms relevant to wheat acreage in that state.
- "Lifetime of a certain type of lightbulb" ⇒ all such lightbulbs produced/available under the target scope.

Survey sampling is a statistical process that involves selecting a sample from a population to conduct a "survey" or "questionnaire". At a basic level, survey sampling involves the following main steps:

- **Sample selection:** choose a sample that is representative of the population via a probabilistic design;
- **Data collection:** measure the characteristic on sampled units (e.g., This involves online surveys, interviews or other methods that give you data on each object/subject in the sample);
- **Estimation:** estimate population parameters (e.g., sample mean, total, proportion, etc.) based on the data;
- **Inference:** results are extrapolated to the population by accounting for the randomness of the sample and quantifying uncertainty due to randomness in sampling (e.g., standard errors, CI).

2 Sampling types and designs

2.1 Probabilistic vs non-probabilistic sampling

- **Probabilistic sampling:** Each population unit has a *known* (and typically nonzero) probability of being selected under a clearly specified sampling design (e.g., SRS, stratified, cluster). Because the randomness comes from the *sampling mechanism*, we can treat estimators (like $\bar{X}$) as random variables and compute their **sampling distributions**, **standard errors**, and **confidence intervals**. In short: *known selection probabilities ⇒ measurable uncertainty.*
- **Non-probabilistic sampling:** The sample is obtained by convenience, self-selection, or volunteer response (e.g., an online poll or an in-class show of hands). Here, selection probabilities are *unknown* and can depend on the outcome itself (e.g., people with stronger opinions respond more). The main risk is **sampling bias**: the sample may not represent the target population, so $\mathbb{E}[\bar{X}] \neq \mu$ and reported "SE/CI" can be meaningless because they ignore selection bias.

Example

Suppose we want the mean height of all students at a university:

- If we take an SRS from the full student roster, then $\bar{X}$ is unbiased for μ and has a well-defined standard error (SE).
- If we sample students who happen to be in the gym at 6pm, the sample is "random" among gym-goers but not among all students. Then $\bar{X}$ may be systematically higher than μ (selection bias), and increasing n does not fix the bias.

2.2 Common probabilistic designs (Probabilistic sampling)

- **Simple Random Sampling (SRS).** Every subset of n units has the same probability of being selected. SRS is the conceptual baseline for survey sampling: estimators are unbiased and variance formulas are simple. However, it may be inefficient or impractical when the population is large, geographically dispersed, or contains important subgroups with very different characteristics.
- **Stratified sampling.** The population is partitioned into *homogeneous strata* (e.g., age groups, regions, gender), and a random sample is taken independently within each stratum (often SRS). This design guarantees representation of all strata and often *reduces variance* compared to SRS, especially when variability within strata is small relative to variability between strata. Correct analysis must account for stratum sizes and sampling fractions.
- **Cluster sampling.** The population is divided into *clusters* (e.g., schools, households, city blocks), which are natural groupings that are typically heterogeneous internally. A random sample of clusters is selected, and all units within chosen clusters are observed. This design is cost-efficient for data collection, but observations within clusters tend to be correlated, which usually *increases variance* relative to SRS and must be reflected in the standard error.
- **Multistage sampling.** Sampling is conducted in multiple stages, combining several designs. For example, one may first sample clusters, then sample units within selected clusters. This approach is common in large-scale surveys where a single-stage design is infeasible. Variance estimation must account for randomness introduced at *each stage*.
- **Systematic random sampling.** A random starting point is chosen, and then every k-th unit is selected from an ordered list. Systematic sampling is simple to implement and often spreads the sample evenly across the population. It behaves similarly to SRS if the ordering is unrelated to the characteristic of interest, but can perform poorly if there is hidden periodicity aligned with the sampling interval.

3 Population parameters vs. sample statistics

A central distinction in survey sampling is between *population parameters* and *sample statistics*. Understanding this distinction is essential for interpreting uncertainty in survey-based inference.

3.1 Population parameters

A **population parameter** is a *fixed (but usually unknown)* numerical value that describes how a characteristic of interest varies across the entire population. It is not random; randomness arises

only because we do not observe the full population. Examples of population parameters **are Mean (average), Median, Variance or standard deviation, and Proportion**.

Example

The *average height of all NCSU freshmen in 2022* is a population parameter. Here,

- the *population* is all NCSU freshmen in 2022, and
- the *feature of interest* is the average height.

If every freshman were measured, this quantity could be computed exactly.

3.2 Sample statistics and estimators

A **sample statistic** is a numerical quantity computed from observed sample data and used to estimate a population parameter. Before data are collected, the sample itself is *random* because it is obtained through a probabilistic sampling procedure (e.g., SRS). Consequently, any statistic computed from the sample is also a *random variable*. In this context, a sample statistic viewed as a random variable is called an **estimator**.

- Population parameter: fixed but unknown
- Sample statistic (estimator): random before sampling, observed after data collection

Example

Let μ denote the true average height of all NCSU freshmen in 2022.

- The population mean μ is a fixed parameter.
- The sample mean $\bar{X}$ is a *random variable* before sampling because different samples give different values.
- Once a particular sample is observed, $\bar{x}$ is a realized numerical value.

Because sample statistics are random, their behavior must be described probabilistically. This leads to the concept of a **sampling distribution**, which quantifies how a statistic (such as $\bar{X}$) varies over repeated samples drawn under the same design. In this chapter, we focus on how the sampling design affects

- the expectation (bias or unbiasedness) of common estimators, and
- their variability, summarized by the *standard error*.

When the population is finite, the sampling design (with or without replacement) plays a crucial role in determining these properties.

4 Finite population

Throughout this chapter, we assume a *finite population* consisting of N units. Each unit has a fixed value of the characteristic of interest:

$$x_1, x_2, \ldots, x_N.$$

These values are treated as *constants*. All randomness in survey sampling arises from the *sampling process*, not from the population itself. Several key numerical summaries of the population are called **population parameters**. They describe global features of how the characteristic varies across the entire population.

- **Population mean:**
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i,$$
which measures the average level of the characteristic.

- **Population total:**
$$\tau = \sum_{i=1}^{N} x_i = N\mu,$$
which is often the primary target in applications such as economics, agriculture, and more.

- **Population variance:**
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2, \qquad \sigma = \sqrt{\sigma^2},$$
which quantifies how much individual units vary around the population mean.

These parameters are *fixed but unknown*. The goal of survey sampling is to estimate them accurately using only a subset of the population.

4.1 Dichotomous (0–1) characteristics

An important special case arises when the characteristic is binary:

$$x_i \in \{0, 1\},$$

where $x_i = 1$ indicates the presence of a trait and $x_i = 0$ its absence. Examples include voter turnout (yes/no), employment status (employed/unemployed), or possession of a certain attribute. In this case:

- the population mean μ equals the **population proportion**
$$p = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

- the population variance simplifies to
$$\sigma^2 = p(1-p).$$

Dichotomous (0–1) characteristics: $\mu = p$ and $\sigma^2 = p(1-p)$

Suppose the population characteristic is binary:

$$x_i \in \{0, 1\}, \qquad i = 1, \ldots, N,$$

where $x_i = 1$ indicates the presence of the trait and $x_i = 0$ its absence.

Population proportion. The population proportion of ones is defined as

$$p \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i.$$

(1) Population mean. The population mean is

$$\mu \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Comparing with the definition of p, we immediately obtain

$$\mu = p.$$

(2) Population variance. The population variance is

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

Using $\mu = p$ and the identity $x_i^2 = x_i$ for $x_i \in \{0,1\}$,

$$(x_i-p)^2 = x_i^2 - 2px_i + p^2 = x_i - 2px_i + p^2.$$

Therefore,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\big(x_i - 2px_i + p^2\big) = \frac{1}{N}\sum_{i=1}^{N} x_i - 2p\cdot\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} p^2.$$

Substituting $\frac{1}{N}\sum_{i=1}^{N} x_i = p$ and $\frac{1}{N}\sum_{i=1}^{N} p^2 = p^2$, we obtain

$$\sigma^2 = p - 2p^2 + p^2 = p(1-p).$$

Thus, estimating a population proportion is mathematically equivalent to estimating a population mean for 0–1 data.

4.2 Population viewed as a random variable

Although the population values $\{x_1, \ldots, x_N\}$ are fixed, it is often convenient to introduce a *conceptual random variable* to summarize the population. Define a random variable X as the characteristic value of a randomly selected population unit. If the distinct values in the population are

$$\zeta_1, \ldots, \zeta_m \quad \text{with frequencies} \quad n_1, \ldots, n_m,$$

then X has probability mass function

$$\mathbb{P}(X = \zeta_j) = \frac{n_j}{N}, \qquad j = 1, \ldots, m.$$

Under this construction,

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

Derivation: $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$

Let the finite population be $\{x_1, \ldots, x_N\}$ (fixed values). Define the conceptual random variable X as the value of a randomly selected unit. Suppose the distinct population values are

$$\zeta_1, \ldots, \zeta_m \quad \text{with frequencies} \quad n_1, \ldots, n_m, \qquad \sum_{j=1}^{m} n_j = N,$$

so that

$$\mathbb{P}(X = \zeta_j) = \frac{n_j}{N}, \qquad j = 1, \ldots, m.$$

(1) Mean. By the definition of expectation for a discrete random variable,

$$\mathbb{E}[X] = \sum_{j=1}^{m} \zeta_j \, \mathbb{P}(X = \zeta_j) = \sum_{j=1}^{m} \zeta_j \frac{n_j}{N} = \frac{1}{N} \sum_{j=1}^{m} n_j \zeta_j.$$

But the multiset $\{x_1, \ldots, x_N\}$ contains the value ζ_j exactly n_j times, hence

$$\sum_{i=1}^{N} x_i = \sum_{j=1}^{m} n_j \zeta_j.$$

Therefore,

$$\mathbb{E}[X] = \frac{1}{N} \sum_{i=1}^{N} x_i = \mu.$$

(2) Variance. Similarly,

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \sum_{j=1}^{m} (\zeta_j - \mu)^2 \, \mathbb{P}(X = \zeta_j) = \sum_{j=1}^{m} (\zeta_j - \mu)^2 \frac{n_j}{N} = \frac{1}{N} \sum_{j=1}^{m} n_j (\zeta_j - \mu)^2.$$

Again using the fact that ζ_j appears n_j times in the population,

$$\sum_{i=1}^{N} (x_i - \mu)^2 = \sum_{j=1}^{m} n_j (\zeta_j - \mu)^2.$$

Hence,

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

Important interpretation. The random variable X does *not* represent a stochastic population. Rather, it is a bookkeeping device that allows population summaries (mean, variance) to be written using probabilistic notation. When sampling is done *with replacement*, each draw has the same distribution as X. When sampling is done *without replacement*, dependence is introduced across draws, leading to variance reductions captured by the finite population correction.

5 Estimators and the sampling distribution

Before any data are collected, a simple random sample (SRS) of size n can be represented as a collection of random variables

$$X_1, X_2, \ldots, X_n,$$

where each X_i denotes the characteristic value observed on the ith sampled unit. The joint distribution of $(X_1, \ldots, X_n)$ is determined entirely by the *sampling design* (with or without replacement).

Once a sample is observed, the realized values are written as

$$x_1, x_2, \ldots, x_n,$$

but it is crucial to remember that *before sampling*, the X_i are random variables.

An **estimator** is a function of the sampled random variables $(X_1, \ldots, X_n)$ used to estimate a population parameter. Two of the most important estimators in survey sampling are:

- **Sample mean**

  $$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

  which estimates the population mean μ.

- **Estimated total**

  $$T = n\bar{X},$$

  which estimates the population total $\tau = \sum_{i=1}^{N} x_i$.

Both $\bar{X}$ and T are random variables *before* data collection. After the data are observed, their realized values $\bar{x}$ and $\hat{\tau}$ are reported as numerical estimates.

The **sampling distribution** of an estimator is the probability distribution of that estimator over all possible samples that could be drawn under the specified sampling design. For example, the sampling distribution of $\bar{X}$ describes how the sample mean would vary if we repeatedly drew many SRS samples of size n from the same population. Key properties of an estimator are defined in terms of its sampling distribution:

- **Bias:** $\mathbb{E}[\bar{X}] - \mu$, which measures systematic error.
- **Variance:** $\mathrm{Var}(\bar{X})$, which measures random fluctuation due to sampling.
- **Standard error (SE):**

  $$\mathrm{SE}(\bar{X}) = \sqrt{\mathrm{Var}(\bar{X})},$$

  which gives the typical scale of estimation error.

Two samples drawn from the same population using the same design will generally produce different estimates. The sampling distribution quantifies this variability and provides the foundation for

- comparing estimators,
- constructing confidence intervals,
- understanding how sampling design (with vs. without replacement) affects uncertainty.

In the sections that follow, we derive the sampling distributions (or their means and variances) of $\bar{X}$ and T under simple random sampling with replacement and without replacement, highlighting the role of independence, dependence, and the finite population correction.

6 SRS with replacement (WR)

Under **simple random sampling with replacement**, each draw is made from the full population, and previously selected units are returned before the next draw. As a result, the sampled values

$$X_1, X_2, \ldots, X_n$$

are *independent and identically distributed (i.i.d.)* copies of the population random variable X.

Formally,

$$X_1, \ldots, X_n \text{ are i.i.d. with } \quad \mathbb{E}[X_i] = \mu, \qquad \mathrm{Var}(X_i) = \sigma^2.$$

This independence assumption is extremely powerful: it allows us to apply standard results from probability theory (linearity of expectation, variance additivity, CLT) without modification. For this reason, WR sampling serves as a useful *baseline* for understanding more complex designs.

Theorem

(Mean and variance of $\bar{X}$ under SRS with replacement) Let $X_1, \ldots, X_n$ be i.i.d. with

$$\mathbb{E}[X_i] = \mu, \qquad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

Define the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and the total estimator} \quad T = N\bar{X}.$$

Then

$$\mathbb{E}[\bar{X}] = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}, \qquad \mathbb{E}[T] = N\mu = \tau.$$

Derivation.

(Mean of $\bar{X}$). Using linearity of expectation,

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\mu = \mu.$$

(Variance of $\bar{X}$). First pull out the constant:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right), \text{where}$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathrm{Var}(X_i) + 2\sum_{1\le i<j\le n}\mathrm{Cov}(X_i, X_j).$$

Because the draws are independent under sampling with replacement, $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, hence

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathrm{Var}(X_i) = n\sigma^2.$$

Substituting back,

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

(Standard error). By definition,

$$\mathrm{SE}(\bar{X}) = \sqrt{\mathrm{Var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

(Mean of the total estimator). Since $T = N\bar{X}$,

$$\mathbb{E}[T] = \mathbb{E}[N\bar{X}] = N\,\mathbb{E}[\bar{X}] = N\mu = \tau.$$

Because $\mathbb{E}[\bar{X}_n] = \mu$, the sample mean $\bar{X}_n$ is called an **unbiased estimator** of the population mean μ. Unbiasedness means that, over repeated samples drawn under the same design, the estimator does not systematically overestimate or underestimate μ.

The variability of $\bar{X}_n$ across different samples is measured by its **standard error**,

$$\mathrm{SE}(\bar{X}_n) = \sqrt{\mathrm{Var}(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}.$$

The standard error represents the typical magnitude of random sampling error: it decreases as the sample size n increases, but only at rate $1/\sqrt{n}$. The standard error

$$\mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

quantifies the typical magnitude of random sampling error. It shows that:

- increasing the sample size n reduces uncertainty at rate $1/\sqrt{n}$;
- halving the standard error requires quadrupling the sample size;
- population variability σ^2 directly controls estimation difficulty.

6.1 CLT approximation for large n

When the sample size n is large, the Central Limit Theorem (CLT) applies:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

Equivalently,

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

This result states that the sampling distribution of the sample mean becomes approximately normal as n increases, regardless of the shape of the population distribution, provided that the population variance σ^2 is finite. Thus, even when the population distribution is skewed or discrete, the distribution of $\bar{X}$ is nearly normal for sufficiently large samples.

Implications. The CLT has several important practical consequences:

- The sampling distribution of $\bar{X}$ can be approximated by a normal distribution, allowing probability calculations and inference using standard normal tables.
- Approximate confidence intervals for μ can be constructed using $\bar{X} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ (or with σ replaced by S in practice).
- The accuracy of the normal approximation improves as n increases, with faster convergence when the population distribution is less skewed.

For these reasons, the CLT provides the theoretical foundation for normal-based confidence intervals and hypothesis tests in survey sampling.

6.2 Unknown σ: plug-in via sample variance

In practice, the population variance σ^2 is typically unknown. It is therefore replaced by the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \text{where} \quad S = \sqrt{S^2}.$$

The denominator $n-1$ ensures that S^2 is an unbiased estimator of σ^2 under i.i.d. sampling. Substituting S for σ yields the *studentized* statistic

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Although S is random, large-sample theory implies that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1) \qquad \text{as } n \to \infty.$$

Implication. This result justifies replacing the unknown standard error $\sigma/\sqrt{n}$ with the estimated standard error $S/\sqrt{n}$ when constructing normal-based confidence intervals and hypothesis tests. For sufficiently large samples, the additional uncertainty introduced by estimating σ is negligible.

Using the sampling distribution to assess accuracy. Knowledge of the (approximate) sampling distribution of $\bar{X}_n$ allows us to evaluate probability statements involving the estimator. In particular, it enables us to quantify the accuracy of using $\bar{X}_n$ to estimate the population mean μ. For example, consider the probability that the estimation error is smaller than a prescribed tolerance $\delta > 0$:

$$\mathbb{P}\big(|\bar{X}_n - \mu| < \delta\big)\,.$$

Under the CLT and using the studentized approximation,

$$\mathbb{P}\big(|\bar{X}_n - \mu| < \delta\big) \approx \mathbb{P}\left(|Z| < \frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - 1,$$

where $Z \sim N(0,1)$ and Φ denotes the standard normal CDF. In practice, σ is replaced by S, yielding a fully data-based approximation. This calculation shows explicitly how estimation accuracy depends on:

- the population variability σ (or its estimate S),
- the sample size n, and
- the desired error tolerance δ.

In particular, increasing n increases the probability that $\bar{X}_n$ lies within δ of μ, but only at the $\sqrt{n}$ rate dictated by the standard error.

6.3 CLT-based confidence interval for μ

A **confidence interval of level** $1-\alpha$ for the parameter μ is a random interval (L, U), constructed from the sample, such that

$$\mathbb{P}(\mu \in (L, U)) \approx 1 - \alpha.$$

The probability is taken with respect to the randomness induced by the sampling procedure. To derive such an interval for μ, we use the Central Limit Theorem. When the sample size n is large,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1).$$

Since σ is unknown in practice, we replace it with the sample standard deviation S, yielding the studentized statistic

$$Z_n = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

which is approximately standard normal for large n.

Let $z_{1-\alpha/2}$ denote the $(1-\alpha/2)$ quantile of the standard normal distribution, so that

$$\mathbb{P}(Z \le z_{1-\alpha/2}) = 1 - \alpha/2, \qquad \mathbb{P}(Z \ge -z_{1-\alpha/2}) = 1 - \alpha/2,$$

for $Z \sim N(0, 1)$. Equivalently,

$$\mathbb{P}\big(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}\big) = 1 - \alpha.$$

Applying this probability statement to Z_n, we obtain the approximation

$$\mathbb{P}\left(-z_{1-\alpha/2} \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

Multiplying all terms by $S/\sqrt{n}$ and rearranging yields

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Therefore, an approximate $100(1-\alpha)\%$ confidence interval for μ is

$$\mu \in \left(\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Interpretation. If the sampling procedure were repeated many times under the same design, and a confidence interval were constructed from each sample using this method, then approximately $100(1-\alpha)\%$ of those intervals would contain the true population mean μ. The confidence level thus describes the long-run performance of the procedure, not the probability that a single realized interval contains μ.

6.4 Estimating the population variance

Assume that $X_1, \dots, X_n$ are i.i.d. random variables with

$$\mathbb{E}[X_i] = \mu, \qquad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

The population variance σ^2 is unknown and must be estimated from the sample. A natural first attempt is the *naive sample variance*

$$\hat{\sigma}^2_{\text{naive}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Derivation of the bias. To compute $\mathbb{E}[\hat{\sigma}^2_{\text{naive}}]$, we first use the identity

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Taking expectations on both sides gives

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n\,\mathbb{E}\big[(\bar{X} - \mu)^2\big].$$

Since the X_i are i.i.d.,

$$\mathbb{E}\big[(X_i - \mu)^2\big] = \sigma^2, \qquad \mathbb{E}\big[(\bar{X} - \mu)^2\big] = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Therefore,

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2.$$

Dividing by n, we obtain

$$\mathbb{E}\big[\hat{\sigma}^2_{\text{naive}}\big] = \frac{n-1}{n}\,\sigma^2.$$

The naive estimator $\hat{\sigma}^2_{\text{naive}}$ systematically underestimates σ^2 and is therefore a *biased* estimator. The bias arises because the sample mean $\bar{X}$ is itself estimated from the data, which reduces the apparent variability of the observations.

Unbiased estimator. To remove this bias, we define the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Using the previous calculation,

$$\mathbb{E}[S^2] = \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n-1}\,(n-1)\sigma^2 = \sigma^2.$$

Hence, S^2 is an **unbiased estimator** of the population variance σ^2. The adjustment from n to $n-1$ in the denominator is known as *Bessel's correction*. Moreover, one can show that

$$S^2 \xrightarrow{p} \sigma^2 \qquad \text{as } n \to \infty.$$

Thus, S^2 is a *consistent estimator* of σ^2, and by continuity of the square root,

$$S = \sqrt{S^2} \xrightarrow{p} \sigma.$$

Connection to the CLT. The consistency of S explains why it can replace the unknown σ in large-sample inference. In particular, substituting S for σ in the standardized statistic

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

does not affect its limiting distribution, leading to the studentized CLT:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1).$$

7 Simple Random Sampling without Replacement (WOR)

In simple random sampling without replacement (WOR), sampled observations are *not independent*, because once a unit is selected, it cannot be selected again. Nevertheless, the design remains symmetric: every population unit has the same probability of being included in the sample. This symmetry ensures unbiased estimation of population quantities, while the dependence structure affects the variance of estimators.

7.1 Dependence and negative covariance

Recall that X_i denotes the ith observation in the sample and $X_1, \ldots, X_n$ are the random variables induced by the sampling design. When sampling is performed *without replacement*, the random variables $X_1, \ldots, X_n$ are *dependent*, because selecting one population unit changes the set of remaining units. It is nevertheless true that each X_i has the same marginal distribution as a randomly selected population element. That is,

$$X_i \sim X,$$

where X denotes the characteristic of a uniformly randomly chosen population member. This follows from the symmetry of simple random sampling: each population unit has the same probability of appearing in any given sample position.

However, dependence becomes evident when we consider conditional distributions. Suppose that $X_{i'} = \zeta_1$ for some $i' \neq i$. Then the conditional distribution of X_i is no longer the same as the population distribution. Instead, it corresponds to sampling uniformly from the remaining $N-1$ population units, with the value ζ_1 removed. In particular,

$$\mathbb{P}(X_i = \zeta_j \mid X_{i'} = \zeta_1) = \frac{n_j - \mathbf{1}\{\zeta_j = \zeta_1\}}{N-1},$$

where n_j is the population frequency of value ζ_j. This conditional distribution differs from the marginal distribution $\mathbb{P}(X = \zeta_j) = n_j/N$.

As a consequence, for $i \neq j$,

$$\mathrm{Cov}(X_i, X_j) < 0.$$

Intuitively, if one sampled unit takes an unusually large value, the remaining population contains slightly fewer large values, making it less likely that another sampled unit is also large. This *negative dependence* is the fundamental reason why estimators under WOR sampling have smaller variance than under sampling with replacement. This reduction in variance is quantified by the finite population correction (FPC), which appears explicitly in the variance of the sample mean.

7.2 Mean and variance of the sample mean

Theorem

Mean and variance of $\bar{X}$ under SRS without replacement. Let an SRS of size n be drawn *without replacement* from a finite population $\{x_1, \ldots, x_N\}$ with population mean

$$\mu = \frac{1}{N} \sum_{k=1}^{N} x_k$$

and population variance

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{N} (x_k - \mu)^2.$$

Let $X_1, \ldots, X_n$ be the sampled values and

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

1) Mean of a single draw. By symmetry of SRS, X_i is equally likely to be any population value $x_1, \ldots, x_N$. Hence, for each i,

$$\mathbb{P}(X_i = x_k) = \frac{1}{N} \quad (k = 1, \ldots, N), \qquad \Rightarrow \qquad \mathbb{E}[X_i] = \sum_{k=1}^{N} x_k \frac{1}{N} = \mu.$$

2) Unbiasedness of the sample mean. Using linearity of expectation,

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu.$$

Therefore, $\bar{X}$ is an **unbiased** estimator of μ:

$$\mathbb{E}[\bar{X}] = \mu.$$

3) Variance decomposition for $\bar{X}$. Start from the general identity

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \left(\sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j)\right).$$

4) Compute $\mathrm{Var}(X_i)$. Since each X_i has the same marginal distribution as a uniformly chosen population element,

$$\mathrm{Var}(X_i) = \sigma^2 \qquad (i = 1, \ldots, n).$$

5) Compute $\mathrm{Cov}(X_i, X_j)$ for $i \neq j$. Fix $i \neq j$. Under WOR sampling, conditional on $X_i = x_k$, the remaining draw X_j is uniformly distributed over the remaining $N - 1$ values:

$$\mathbb{P}(X_j = x_\ell \mid X_i = x_k) = \frac{1}{N-1} \quad (\ell \neq k).$$

Hence,

$$\mathbb{E}[X_j \mid X_i = x_k] = \frac{1}{N-1}\sum_{\ell \neq k} x_\ell = \frac{N\mu - x_k}{N-1}.$$

Now compute $\mathbb{E}[X_i X_j]$ using iterated expectation:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[\, X_i\, \mathbb{E}[X_j \mid X_i]\,] = \mathbb{E}\left[X_i \cdot \frac{N\mu - X_i}{N-1}\right] = \frac{1}{N-1}\left(N\mu\,\mathbb{E}[X_i] - \mathbb{E}[X_i^2]\right).$$

Using $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) + (\mathbb{E}[X_i])^2 = \sigma^2 + \mu^2$,

$$\mathbb{E}[X_i X_j] = \frac{1}{N-1}\left(N\mu^2 - (\sigma^2 + \mu^2)\right) = \frac{(N-1)\mu^2 - \sigma^2}{N-1} = \mu^2 - \frac{\sigma^2}{N-1}.$$

Therefore,

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = \left(\mu^2 - \frac{\sigma^2}{N-1}\right) - \mu^2 = -\frac{\sigma^2}{N-1}.$$

This is negative, capturing the **negative dependence** under WOR sampling.

6) Put everything together. Plug $\mathrm{Var}(X_i) = \sigma^2$ and $\mathrm{Cov}(X_i, X_j) = -\sigma^2/(N-1)$ into Step 3:

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\left(n\sigma^2 + 2\binom{n}{2}\left(-\frac{\sigma^2}{N-1}\right)\right) = \frac{1}{n^2}\left(n\sigma^2 - \frac{n(n-1)\sigma^2}{N-1}\right).$$

Factor and simplify:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n^2}\, n\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

Conclusion.

$$\mathbb{E}[\bar{X}] = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

The factor

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

is called the **finite population correction (FPC)**.

Interpretation. The sample mean under WOR sampling remains unbiased for μ, but its variance is smaller than under sampling with replacement, where $\mathrm{Var}(\bar{X}) = \sigma^2/n$. The reduction factor $\frac{N-n}{N-1}$ reflects the negative dependence introduced by sampling without replacement. In particular, the difference between sampling with and without replacement depends on the *sampling fraction* $f = n/N$:

- If $n \ll N$ (very small sampling fraction), then

$$\frac{N-n}{N-1} \approx 1,$$

and the variance under WOR is essentially the same as under WR. In this regime, treating the sample as i.i.d. is a good approximation.

- If n is a non-negligible fraction of N, the finite population correction can be substantial. Ignoring WOR and using WR formulas then *overestimates* the true uncertainty, leading to overly conservative standard errors and confidence intervals.

Thus, whether "with replacement" matters is determined not by the sample size alone, but by how large the sample is relative to the population.

7.3 When does WOR matter?

Let

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

denote the **sampling fraction**.

- If f is very small (the population is much larger than the sample), then $\frac{N-n}{N-1} \approx 1$, and formulas derived under sampling with replacement provide good approximations.
- If f is not negligible, ignoring the FPC leads to *overestimation* of the variance and standard error.

Example

Numerical illustration of the FPC. If $N = 1000$ and $n = 50$,

$$\frac{N-n}{N-1} = \frac{950}{999} \approx 0.951.$$

If instead $N = 200$ and $n = 50$,

$$\frac{150}{199} \approx 0.754,$$

indicating a much stronger variance reduction.

7.4 Estimating standard errors under WOR

In practice, the population variance σ^2 is unknown and must be estimated from the sample.

Estimating the population variance

Under WOR sampling, the usual sample variance with denominator $(n-1)$ is not exactly unbiased for

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

Rice shows that an unbiased estimator is

$$\widehat{\sigma^2} = \left(1 - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

When N is large, the factor $(1 - 1/N)$ is close to 1, and the usual sample variance provides an excellent approximation.

Estimated variance and standard error of $\bar{X}$

Replacing σ^2 by the sample variance and using the approximate FPC, a practical estimator of the variance of $\bar{X}$ is

$$\widehat{\mathrm{Var}}(\bar{X}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right), \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Thus, the estimated standard error is

$$\widehat{\mathrm{SE}}(\bar{X}) = s_{\bar{X}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}.$$

Estimating a population total

For the total estimator $T = N\bar{X}$,

$$\mathrm{Var}(T) = N^2\,\mathrm{Var}(\bar{X}), \qquad \widehat{\mathrm{SE}}(T) = N\,s_{\bar{X}}.$$

Dichotomous case: estimating a proportion

If $X_i \in \{0, 1\}$ indicates the presence of a characteristic, then $\mu = p$ is the population proportion and

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A commonly used estimated variance under WOR is

$$\widehat{\mathrm{Var}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right), \qquad \widehat{\mathrm{SE}}(\hat{p}) = \sqrt{\widehat{\mathrm{Var}}(\hat{p})}.$$

7.5 Normal approximation and confidence intervals

Although the X_i are dependent under WOR sampling, finite-population central limit theorems imply that, when n is large and still small relative to N, the sample mean is approximately normal:

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \approx N(0, 1), \qquad \sigma_{\bar{X}} = \sqrt{\mathrm{Var}(\bar{X})}.$$

In practice, $\sigma_{\bar{X}}$ is replaced by $s_{\bar{X}}$. An approximate $100(1-\alpha)\%$ confidence interval for μ is therefore

$$\mu \in \left(\bar{X} - z_{1-\alpha/2}\,s_{\bar{X}},\ \bar{X} + z_{1-\alpha/2}\,s_{\bar{X}}\right), \qquad s_{\bar{X}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}.$$

Example

Interpreting a standard error. If $\bar{X}$ is approximately normal, then

$$\mathbb{P}\left(|\bar{X} - \mu| \le 2\,\sigma_{\bar{X}}\right) \approx 0.95.$$

Thus, the estimated standard error provides a direct scale for typical estimation error: errors of one to two standard errors are common, while much larger errors are unlikely.

This is a personal study purposed notes, based on a lecture slide given by Prof. Ana-Maria Staicu in ST 502, NC state university and Rice (3rd ed.) Chapter 5–6.