Statistical Inference

Chapter 1: Sampling, Sampling Distributions and Experimental Designs

Abdelhakim Necir

Master: Probability, Statistics and Applications

Academic Year 2026–2027

Chapter Outline

1. Basic Concepts
2. Sampling and Data Collection
3. Experimental Designs
4. Sampling Distributions
5. Central Limit Theorem
6. Summary

Population and Sample

Population

A population is the complete collection of individuals, objects, measurements, or outcomes about which statistical information is required.

Sample

A sample is a collection of observations selected from a population.

If the sample contains n observations, we write

$$X_1, \ldots, X_n.$$

The integer n is called the sample size.

Random Sample

A common framework in statistical inference is

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P.$$

If the distribution depends on an unknown parameter θ, we write

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta.$$

Important

The i.i.d. assumption means that the observations are independent and have the same distribution.

Parameter and Statistic

Parameter

A parameter is a numerical characteristic of a population.

Examples:

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

Statistic

A statistic is a measurable function of the sample which does not depend on unknown parameters.

Examples:

$$\overline{X}, \qquad S^2, \qquad \widehat{p}.$$

Examples of Statistics

The sample mean is

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

The sample variance is

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

For Bernoulli observations, the sample proportion is

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Why Do We Use a Sample?

A complete study of the whole population is called a census. Sampling is often preferred because:

- it is less expensive;
- it is faster;
- the population may be very large or infinite;
- some measurements may be destructive;
- a good sample can provide accurate information.

Methods of Data Collection

Observational Study

The investigator observes without modifying the conditions.

Survey

Information is collected using questionnaires, interviews, or forms.

Experiment

Treatments are deliberately applied and responses are observed.

Types of Sampling

Sampling methods can be divided into:

- probability sampling;
- non-probability sampling.

Probability Sampling

Each population unit has a known probability of being selected.

Non-Probability Sampling

Selection probabilities are not controlled by a random mechanism.

Probability Sampling Methods

- **Simple random sampling:** every sample of size n has the same probability.
- **Systematic sampling:** units are selected using a fixed interval.
- **Stratified sampling:** the population is divided into strata.
- **Cluster sampling:** groups or clusters are selected.

Remark

Probability sampling is generally preferred for formal statistical inference.

Basic Concepts in Experimental Design

- **Experimental unit:** the unit receiving a treatment.
- **Factor:** explanatory variable studied in the experiment.
- **Level:** possible value of a factor.
- **Treatment:** experimental condition applied to a unit.
- **Response variable:** measured outcome.

Principles of Experimental Design

Randomization

Treatments are assigned by a random mechanism.

Replication

Each treatment is applied to several experimental units.

Control

A reference group is used for comparison.

Sampling Distribution

A statistic is a random variable because its value changes from one sample to another.

Definition

The probability distribution of a statistic is called its sampling distribution.

Sampling distributions are the basis of:

- point estimation;
- confidence intervals;
- hypothesis testing.

Sampling Distribution of the Sample Mean

Let $X_1, \ldots, X_n$ be i.i.d. random variables such that

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

The sample mean is

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

Then

$$\mathbb{E}(\overline{X}) = \mu, \qquad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Standard Error of the Sample Mean

Since

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n},$$

the standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$. This quantity is called the standard error of the sample mean:

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

Consequence

Increasing the sample size reduces the standard error.

Normal Population

If

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

then

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Equivalently,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Important

This result is exact when the population is normal.

Sampling Distribution of a Proportion

Let $X_i \sim \text{Bernoulli}(p)$, $i = 1, \ldots, n$. The sample proportion is

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then

$$\mathbb{E}(\widehat{p}) = p, \qquad \text{Var}(\widehat{p}) = \frac{p(1-p)}{n}.$$

Normal Approximation for a Proportion

For large n,

$$\widehat{p} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Thus,

$$\frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \approx \mathcal{N}(0,1).$$

A common rule of thumb is

$$np \geq 5 \qquad \text{and} \qquad n(1-p) \geq 5.$$

Difference of Two Sample Means

Let $X_1, \ldots, X_{n_1}$ be a sample from a population with mean μ_1 and variance σ_1^2. Let $Y_1, \ldots, Y_{n_2}$ be an independent sample from a population with mean μ_2 and variance σ_2^2. Then

$$\mathbb{E}(\overline{X} - \overline{Y}) = \mu_1 - \mu_2.$$

Variance of the Difference

Because the samples are independent,

$$\mathrm{Var}(\overline{X} - \overline{Y}) = \mathrm{Var}(\overline{X}) + \mathrm{Var}(\overline{Y}).$$

Therefore,

$$\mathrm{Var}(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

If both populations are normal, then

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Sampling Distribution of the Sample Variance

Assume

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

The sample variance is

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

Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Important

This exact result requires normality of the population.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. random variables such that

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

Then

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

Meaning of the Central Limit Theorem

For large n,

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

The original population does not need to be normal.

Main Idea

For large samples, the sample mean is approximately normally distributed.

Application

Suppose $\mu = 50$, $\sigma = 10$, and $n = 100$. Then

$$\overline{X} \approx \mathcal{N}\left(50, \frac{10^2}{100}\right).$$

Thus,

$$\mathbb{E}(\overline{X}) = 50, \qquad \mathrm{SD}(\overline{X}) = 1.$$

Therefore,

$$P(48 \leq \overline{X} \leq 52) \approx P(-2 \leq Z \leq 2).$$

Summary

In this chapter, we studied:

- population, sample, parameter, and statistic;
- methods of data collection;
- types of sampling;
- basic principles of experimental design;
- sampling distribution of the sample mean;
- sampling distribution of a sample proportion;
- sampling distribution of the difference of two means;
- sampling distribution of the sample variance;
- Central Limit Theorem.

Key Formulas

$$\mathbb{E}(\overline{X}) = \mu, \qquad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

$$\mathbb{E}(\widehat{p}) = p, \qquad \mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}.$$

$$\mathrm{Var}(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

End of Chapter 1

End of Chapter 1 Next chapter: Point Estimation