Statistical Inference

Course Notes

Prof. Abdelhakim Necir

Mohamed Khider University, Biskra
Faculty of Exact Sciences
Department of Mathematics
Algeria

Academic Year 2026–2027

Contents

1 Sampling, Sampling Distributions and Experimental Designs **6**

1.1 Introduction . . . 6
1.2 Population and Sample . . . 6
1.2.1 Population . . . 6
1.2.2 Sample . . . 7
1.3 Parameter and Statistic . . . 7
1.4 Why Do We Use a Sample? . . . 8
1.4.1 Cost . . . 8
1.4.2 Time . . . 8
1.4.3 Destructive Measurements . . . 8
1.4.4 Very Large or Infinite Populations . . . 8
1.4.5 Statistical Efficiency . . . 8
1.5 Methods of Data Collection . . . 9
1.5.1 Observational Studies . . . 9
1.5.2 Surveys . . . 9
1.5.3 Experiments . . . 9
1.6 Types of Sampling . . . 9
1.6.1 Probability Sampling . . . 9
1.6.2 Non-Probability Sampling . . . 10
1.7 Basic Principles of Experimental Design . . . 10
1.7.1 Experimental Unit . . . 10
1.7.2 Factor . . . 10
1.7.3 Treatment . . . 10
1.7.4 Response Variable . . . 10
1.7.5 Randomization . . . 11
1.7.6 Replication . . . 11
1.7.7 Control . . . 11
1.8 Sampling Distributions . . . 11
1.9 Sampling Distribution of the Sample Mean . . . 11
1.9.1 Case of a Normal Population . . . 12
1.10 Sampling Distribution of a Sample Proportion . . . 12
1.11 Sampling Distribution of the Difference of Two Means . . . 13
1.12 Sampling Distribution of the Sample Variance . . . 14
1.13 The Central Limit Theorem . . . 15
1.14 Applications of the Central Limit Theorem . . . 15
1.14.1 Approximation of the Sample Mean . . . 15
1.14.2 Approximation of a Sample Proportion . . . 16
1.15 Standard Errors . . . 16

1.16 Summary . . . 17

2 Point Estimation **18**
2.1 Introduction to Point Estimation . . . 18
2.2 Parametric and Nonparametric Estimation . . . 18
2.2.1 Parametric Estimation . . . 18
2.2.2 Nonparametric Estimation . . . 19
2.3 Point Estimators and Estimates . . . 19
2.4 Random Variables, Estimators, and Unknown Parameters . . . 20
2.5 Unknown Parameters . . . 20
2.6 Methods of Point Estimation . . . 21
2.6.1 An Introductory Example . . . 21
2.6.2 Maximum Likelihood Method . . . 21
2.6.3 The Log-Likelihood Function . . . 22
2.6.4 Likelihood Functions for Common Models . . . 22
2.6.5 Maximum Likelihood Estimator . . . 23
2.6.6 Method of Moments . . . 23
2.7 Properties of Estimators . . . 23
2.7.1 Bias . . . 24
2.7.2 Asymptotic Bias . . . 24
2.7.3 Variance . . . 24
2.7.4 Mean Squared Error . . . 25
2.7.5 Efficiency . . . 25
2.7.6 Admissibility . . . 25
2.7.7 Consistency . . . 25
2.7.8 Convergence in Probability . . . 25
2.7.9 Convergence in Mean of Order k . . . 26
2.8 Sufficiency . . . 26
2.8.1 Formal Definition . . . 26
2.8.2 Fisher–Neyman Factorization Theorem . . . 27
2.8.3 Proof of the Factorization Theorem . . . 27
2.8.4 Examples of Sufficient Statistics . . . 28
2.9 Completeness . . . 29
2.9.1 Formal Definition . . . 29
2.9.2 Proofs of Completeness for Common Models . . . 30
2.9.3 Completeness and Exponential Families . . . 32
2.10 Usual Estimators . . . 33
2.10.1 Quantitative Case . . . 33
2.10.2 Qualitative Case: Unknown Proportion . . . 33
2.11 Fisher Information and Score . . . 33
2.11.1 The Score Function . . . 34
2.11.2 Fisher Information . . . 34
2.11.3 Example: Fisher Information for Exponential Model . . . 35
2.11.4 Efficiency and Fisher Information . . . 35
2.11.5 Fisher Information and Completeness . . . 35
2.12 Cramér–Rao–Fréchet–Darmois Bound . . . 35
2.12.1 Illustrative Example: Normal Mean . . . 36
2.12.2 Minimum Variance Unbiased Estimator (MVUE) . . . 37

2.13 Conclusion . . . 38

3 Confidence Interval Estimation **39**
3.1 Point Estimation versus Confidence Interval Estimation . . . 39
3.2 Confidence Level and Risk Level . . . 40
3.2.1 Interpretation of a Confidence Interval . . . 40
3.3 Construction of Confidence Intervals . . . 41
3.4 Confidence Interval for a Mean . . . 41
3.4.1 Known Variance . . . 42
3.4.2 Unknown Variance . . . 42
3.4.3 Summary for the Mean . . . 43
3.5 Confidence Interval for a Variance . . . 44
3.6 Confidence Interval for a Proportion . . . 45
3.7 Confidence Intervals for Two Populations . . . 45
3.7.1 Difference of Two Means (Known Variances) . . . 46
3.7.2 Difference of Two Means (Unknown Equal Variances) . . . 46
3.7.3 Ratio of Two Variances . . . 47
3.7.4 Difference of Two Proportions . . . 47
3.8 Influence of Sample Size on Precision . . . 47
3.9 Determination of the Optimal Sample Size . . . 48
3.9.1 Sample Size for Estimating a Mean . . . 48
3.9.2 Sample Size for Estimating a Proportion . . . 48
3.9.3 Factors Affecting the Required Sample Size . . . 49
3.10 Conclusion . . . 49

4 Introduction to Statistical Tests **51**
4.1 Objectives of Statistical Tests . . . 51
4.2 Null and Alternative Hypotheses . . . 51
4.2.1 Two-Sided Alternative . . . 52
4.2.2 Right-Sided Alternative . . . 52
4.2.3 Left-Sided Alternative . . . 52
4.3 Types of Statistical Tests . . . 52
4.4 Neyman–Pearson Principle . . . 53
4.5 Critical and Acceptance Regions . . . 53
4.6 Significance Level . . . 54
4.7 Type I and Type II Errors . . . 54
4.8 Power Function of a Test . . . 55
4.9 The p-value Approach . . . 55
4.9.1 Calculation of the p-value . . . 55
4.10 Decision Table . . . 56
4.11 General Steps of a Statistical Test . . . 56
4.12 Standard Statistical Tables . . . 56
4.12.1 Normal Distribution . . . 57
4.12.2 Student's t Distribution . . . 57
4.12.3 Chi-Square Distribution . . . 57
4.12.4 Fisher–Snedecor Distribution . . . 57
4.13 Parametric Tests . . . 57
4.14 Test for a Population Mean . . . 57
4.14.1 Known Variance . . . 58

4.14.2 Unknown Variance . . . 58
4.15 Test for a Population Variance . . . 58
4.16 Test for a Population Proportion . . . 59
4.17 Comparison of Two Population Means . . . 59
4.17.1 Known Variances . . . 59
4.17.2 Unknown Variances . . . 60
4.18 Comparison of Two Population Variances (CORRIGÉ ET ENRICHI) . . . 60
4.19 Comparison of Two Population Proportions . . . 61
4.20 Illustrative Example: Test for a Population Mean . . . 61
4.21 Duality Between Tests and Confidence Intervals . . . 62
4.22 Summary of Parametric Tests (ENRICHI) . . . 63
4.23 Conclusion . . . 63

Chapter 1

Sampling, Sampling Distributions and Experimental Designs

1.1 Introduction

Statistical inference is based on the principle that information obtained from a sample can be used to draw conclusions about a larger population.

In many practical situations, observing the whole population is impossible, too expensive, too time-consuming, or simply unnecessary. Instead, a subset of units is selected, and the information collected from this subset is used to learn about unknown characteristics of the population.

This chapter introduces the basic concepts of sampling and experimental design, and studies the sampling distributions of some commonly used statistics.

The main objectives are:

- to distinguish between a population and a sample;
- to distinguish between a parameter and a statistic;
- to understand the main methods of data collection;
- to describe the principal sampling designs;
- to introduce the basic principles of experimental design;
- to derive the sampling distributions of important statistics;
- to understand the role of the Central Limit Theorem.

1.2 Population and Sample

1.2.1 Population

Definition 1.1. A *population* is the complete collection of individuals, objects, measurements, or outcomes about which statistical information is required.

A population may be finite or infinite.

Examples include:

- all students registered at a university during a given academic year;
- all manufactured components produced by a factory during one month;
- all households in a given region;
- all possible measurements generated by a probability model.

In a probabilistic formulation, the population is described by a probability distribution, often denoted by P_θ, or by a family

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where θ represents an unknown parameter.

1.2.2 Sample

Definition 1.2. 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*.

When the observations are independent and identically distributed according to a common probability distribution P_θ, we write

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

Remark 1.1. The i.i.d. assumption is fundamental in classical statistical inference. However, in other areas such as time series, spatial statistics, survival analysis, and stochastic processes, observations may be dependent.

1.3 Parameter and Statistic

One of the most important distinctions in statistical inference is the difference between a population parameter and a sample statistic.

Definition 1.3. A *parameter* is a numerical characteristic of a population or of its probability distribution.

Typical examples are:

$$\mu = \mathrm{E}(X),$$

the population mean,

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

the population variance, and

$$p,$$

a population proportion.

Parameters are generally unknown.

Definition 1.4. A *statistic* is any measurable function of the observed sample that does not depend on unknown parameters.

For example, the sample mean is

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

the sample variance is

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

and, for Bernoulli observations, the sample proportion is

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

Remark 1.2. A parameter describes the population, whereas a statistic is computed from the sample.

1.4 Why Do We Use a Sample?

A complete investigation of all units in a population is called a *census*.

Although a census may appear ideal, it is often impractical.

Sampling is commonly preferred for several reasons.

1.4.1 Cost

Collecting information from every member of a large population may require substantial financial resources.

1.4.2 Time

A census may require a long period of data collection, whereas a carefully designed sample can provide useful information much more rapidly.

1.4.3 Destructive Measurements

Some measurements destroy or alter the object being studied.

For example, a component tested until failure cannot be returned to the production process.

1.4.4 Very Large or Infinite Populations

In some situations, the population is extremely large or naturally modeled as infinite.

A complete census is then impossible.

1.4.5 Statistical Efficiency

A properly designed sample can provide sufficiently accurate information with substantially fewer observations than a complete census.

1.5 Methods of Data Collection

Statistical data can be obtained in different ways.

1.5.1 Observational Studies

In an observational study, the investigator records information without deliberately modifying the conditions of the individuals or experimental units.

Observational studies are useful when intervention is impossible, impractical, or ethically inappropriate.

1.5.2 Surveys

A survey collects information directly from individuals, usually through questionnaires, interviews, or electronic forms.

The reliability of the results depends strongly on the sampling design, response rate, and formulation of the questions.

1.5.3 Experiments

In an experiment, the investigator deliberately applies one or more conditions, called *treatments*, and observes their effects on a response variable.

Experiments are especially useful when the objective is to study causal relationships.

Remark 1.3. A statistical analysis cannot compensate for a poor data collection process. The quality of the conclusions depends on both the statistical method and the design by which the data were obtained.

1.6 Types of Sampling

Sampling methods can be broadly divided into probability sampling and non-probability sampling.

1.6.1 Probability Sampling

In probability sampling, each population unit has a known probability of being selected.

This makes it possible to quantify sampling variability.

Simple Random Sampling

Definition 1.5. A *simple random sample* of size n is selected in such a way that every possible sample of size n has the same probability of being chosen.

Simple random sampling is the basic model underlying much of classical statistical inference.

Systematic Sampling

In systematic sampling, a starting unit is selected and then units are chosen according to a fixed interval.

For example, one may select every k-th unit after a random starting point.

Stratified Sampling

The population is divided into subpopulations called *strata*.

A sample is then selected from each stratum.

Stratification is useful when the population contains important subgroups that should all be represented in the sample.

Cluster Sampling

The population is divided into natural groups called *clusters*.

A number of clusters are selected, and observations are collected from units belonging to the selected clusters.

Cluster sampling is often useful when the population is geographically dispersed.

1.6.2 Non-Probability Sampling

In non-probability sampling, the selection probabilities are not controlled by a random mechanism.

Convenience Sampling

A convenience sample consists of observations that are easy to obtain.

Such samples may be strongly biased and are generally unsuitable for formal statistical inference about a population.

Remark 1.4. Probability sampling methods are generally preferred when the objective is to make inferential statements about a population.

1.7 Basic Principles of Experimental Design

An experiment is designed to study the effect of one or more factors on a response variable.

Several concepts are fundamental.

1.7.1 Experimental Unit

An *experimental unit* is the smallest unit to which a treatment is applied.

1.7.2 Factor

A *factor* is an explanatory variable whose effect on the response is being investigated.

The possible values or categories of a factor are called its *levels*.

1.7.3 Treatment

A *treatment* is a specific experimental condition applied to an experimental unit.

1.7.4 Response Variable

The *response variable* is the outcome measured after applying a treatment.

1.7.5 Randomization

Randomization consists of assigning treatments to experimental units by a random mechanism.

Its main purpose is to reduce systematic bias and to help create comparable groups.

1.7.6 Replication

Replication means applying each treatment to several experimental units.

Replication makes it possible to estimate natural variability and generally improves the precision of statistical conclusions.

1.7.7 Control

A control group provides a reference condition against which the effect of a treatment can be compared.

Remark 1.5. Randomization, replication, and control are three basic principles of sound experimental design.

1.8 Sampling Distributions

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

Definition 1.6. The probability distribution of a statistic is called its *sampling distribution.*

Sampling distributions are fundamental because they describe the variability that results from random sampling.

They provide the theoretical basis for estimation, confidence intervals, and hypothesis testing.

1.9 Sampling Distribution of the Sample Mean

Let

$$X_1, \ldots, X_n$$

be independent and identically distributed random variables such that

$$\mathrm{E}(X_i) = \mu$$

and

$$\mathrm{Var}(X_i) = \sigma^2.$$

The sample mean is defined by

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

Its expectation is

$$\mathrm{E}(\overline{X}) = \mathrm{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right).$$

Using linearity of expectation,

$$\mathrm{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \mu.$$

Therefore,

$$\mathrm{E}(\overline{X}) = \mu.$$

Its variance is

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right).$$

By independence,

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

Hence,

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

The standard deviation of the sampling distribution of $\overline{X}$ is

$$\frac{\sigma}{\sqrt{n}}.$$

This quantity is called the *standard error of the sample mean.*

1.9.1 Case of a 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,

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

This result is exact for every sample size n when the parent population is normal.

1.10 Sampling Distribution of a Sample Proportion

Suppose that

$$X_1, \ldots, X_n$$

are independent Bernoulli random variables with

$$P_\theta(X_i = 1) = p$$

and

$$P_\theta(X_i = 0) = 1 - p.$$

Then

$$X_i \sim \text{Bernoulli}(p).$$

The sample proportion is

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

Since

$$\mathrm{E}(X_i) = p,$$

we obtain

$$\mathrm{E}(\widehat{p}) = p.$$

Also,

$$\mathrm{Var}(X_i) = p(1-p),$$

so

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

For sufficiently large n, the sampling distribution of $\widehat{p}$ is approximately normal:

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

Therefore,

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

A commonly used practical rule for the normal approximation is that both np and $n(1-p)$ should be sufficiently large.

A frequent rule of thumb is

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

Remark 1.6. The conditions above are practical guidelines rather than exact mathematical requirements. Some textbooks use the more conservative threshold 10.

1.11 Sampling Distribution of the Difference of Two Means

Consider two independent populations.

Let

$$X_1, \ldots, X_{n_1}$$

be an i.i.d. sample from a population with mean μ_1 and variance σ_1^2, and let

$$Y_1, \ldots, Y_{n_2}$$

be an independent i.i.d. sample from a second population with mean μ_2 and variance σ_2^2.

Define

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$$

and

$$\overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j.$$

By linearity of expectation,

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

Because the two samples are independent,

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

If both populations are normal, then

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

Consequently,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1).$$

1.11.1 Case of Unknown but Equal Variances

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown), we use the *pooled variance estimator*:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The test statistic is then

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

This result will be used in Chapter 4 for comparing two population means.

1.12 Sampling Distribution of the Sample Variance

Let

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

The sample variance is

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

The following result is fundamental.

Theorem 1.1. *If*

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

then

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

The number $n-1$ is called the number of *degrees of freedom*.

Theorem 1.2 (Independence of $\overline{X}$ and S^2)**.** *If*

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

then $\overline{X}$ *and* S^2 *are independent random variables.*

This independence result is crucial for deriving the Student t-distribution.

Remark 1.7. The chi-square distribution above is an exact result under the assumption that the parent population is normal.

This result will later be used to construct confidence intervals and hypothesis tests concerning the population variance.

1.12.1 The Student t-Distribution

The Student t-distribution is one of the most important distributions in statistical inference, particularly when the population variance is unknown.

Definition 1.7 (Student t-Distribution)**.** Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_\nu$ be independent random variables. Then the random variable

$$T = \frac{Z}{\sqrt{V/\nu}}$$

follows a *Student t-distribution* with ν degrees of freedom, denoted by $T \sim t_\nu$.

The probability density function of the t-distribution with ν degrees of freedom is

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \qquad t \in \mathbb{R}.$$

Properties of the t-Distribution

- The t-distribution is symmetric about zero.
- $\mathrm{E}(T) = 0$ for $\nu > 0$.
- $\mathrm{Var}(T) = \frac{\nu}{\nu-2}$ for $\nu > 2$.
- As $\nu \to \infty$, the t-distribution converges to the standard normal distribution: $t_\nu \to \mathcal{N}(0,1)$.

Application to the Sample Mean with Unknown Variance

When the population variance σ^2 is unknown, we replace it by the sample variance S^2.

Theorem 1.3. *If*

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

then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Proof. We can write

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{S/\sigma} = \frac{Z}{\sqrt{V/(n-1)}},$$

where

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

and

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

By the independence theorem, Z and V are independent. Therefore, by definition, $T \sim t_{n-1}$. □

This result will be used extensively in Chapters 3 and 4 for confidence intervals and hypothesis tests concerning the population mean when the variance is unknown.

1.12.2 The Fisher F-Distribution

The Fisher F-distribution is used primarily for comparing two population variances.

Definition 1.8 (Fisher F-Distribution)**.** Let $V_1 \sim \chi^2_{\nu_1}$ and $V_2 \sim \chi^2_{\nu_2}$ be independent random variables. Then the random variable

$$F = \frac{V_1/\nu_1}{V_2/\nu_2}$$

follows a *Fisher F-distribution* with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom, denoted by $F \sim F_{\nu_1,\nu_2}$.

Properties of the F-Distribution

- The F-distribution is positively skewed and takes only positive values: $F > 0$.
- $\mathrm{E}(F) = \frac{\nu_2}{\nu_2 - 2}$ for $\nu_2 > 2$.
- If $F \sim F_{\nu_1,\nu_2}$, then $1/F \sim F_{\nu_2,\nu_1}$.

Application to the Ratio of Two Variances

Theorem 1.4. *Let*

$$X_1, \ldots, X_{n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$$

and

$$Y_1, \ldots, Y_{n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$$

be two independent samples. Then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

In particular, under the null hypothesis $\sigma_1^2 = \sigma_2^2$, we have

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}.$$

This result will be used in Chapters 3 and 4 for confidence intervals and hypothesis tests concerning the ratio of two population variances.

1.13 The Central Limit Theorem

The Central Limit Theorem is one of the most important results in probability and statistics.

It explains why the normal distribution appears so frequently in statistical inference.

Theorem 1.5 (Central Limit Theorem)**.** *Let*

$$X_1, X_2, \ldots$$

be independent and identically distributed random variables such that

$$\mathrm{E}(X_i) = \mu$$

and

$$\mathrm{Var}(X_i) = \sigma^2 < \infty.$$

Then, as $n \to \infty$*:*

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \to \mathcal{N}(0,1) \quad \textit{(in distribution).}$$

Equivalently,

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

for sufficiently large n.

The importance of this theorem is that the original population does not need to be normally distributed.

The approximation improves as the sample size increases, although the rate of improvement depends on the shape of the underlying distribution.

Remark 1.8. A common practical rule is that the normal approximation is adequate when $n \geq 30$. However, for highly skewed or heavy-tailed distributions, larger sample sizes may be required.

1.14 Applications of the Central Limit Theorem

1.14.1 Approximation of the Sample Mean

For large n,

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

This approximation allows probabilities involving the sample mean to be evaluated using the standard normal distribution.

Example 1.1. Suppose a population has mean $\mu = 50$ and standard deviation $\sigma = 10$.

For a random sample of size $n = 100$, the sample mean has approximately the distribution

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

Thus,

$$\mathrm{E}(\overline{X}) = 50$$

and

$$\mathrm{SD}(\overline{X}) = 1.$$

For example,

$$P(48 \leq \overline{X} \leq 52)$$

can be written as

$$P\left(-2 \leq \frac{\overline{X} - 50}{1} \leq 2\right),$$

and therefore

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

where $Z \sim \mathcal{N}(0, 1)$.

From the standard normal table,

$$P(-2 \leq Z \leq 2) \approx 0.9545.$$

1.14.2 Approximation of a Sample Proportion

Since the sample proportion is the mean of Bernoulli random variables, the Central Limit Theorem gives

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

for sufficiently large sample size.

This approximation will later be used in the construction of confidence intervals and hypothesis tests concerning population proportions.

1.15 Standard Errors

The standard deviation of the sampling distribution of a statistic is called its *standard error*.

For the sample mean,

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

For the sample proportion,

$$\mathrm{SE}(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

For the difference between two independent sample means,

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

When the population quantities involved are unknown, the standard error is usually estimated from the sample.

An important consequence is that increasing the sample size reduces the standard error.

For example,

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

so multiplying the sample size by 4 divides the standard error by 2.

Thus, larger samples generally provide more precise statistical information.

1.16 Summary of Key Sampling Distributions

The following list summarizes the main sampling distributions introduced in this chapter:

- **Sample Mean:** $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ (Normal population or large n)
- **Z-statistic:** $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ (σ^2 known)
- **T-statistic:** $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$ (σ^2 unknown, normal population)
- **Chi-square statistic:** $Q = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (Normal population)
- **F-statistic:** $F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$ (Two independent normal populations)
- **Sample Proportion:** $\widehat{p} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ (Large sample: $np \geq 5, n(1-p) \geq 5$)

These distributions form the foundation for the confidence intervals and hypothesis tests developed in Chapters 3 and 4.

1.17 Summary

The central framework of this chapter is based on the distinction between population quantities and sample quantities.

Population characteristics include parameters such as

$$\mu, \qquad \sigma^2, \qquad p,$$

whereas samples produce statistics such as

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

Sampling provides a practical way to obtain information about a population, provided that the sampling design is appropriate.

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

For the sample mean,

$$\mathrm{E}(\overline{X}) = \mu,$$

and

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

For a sample proportion,

$$\mathrm{E}(\widehat{p}) = p,$$

and

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

For two independent sample means,

$$\mathrm{E}(\overline{X} - \overline{Y}) = \mu_1 - \mu_2,$$

and

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

For a normal population,

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

The Student t-distribution arises when estimating the mean with unknown variance:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

The Fisher F-distribution is used for comparing two variances:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}.$$

Finally, the Central Limit Theorem provides the asymptotic approximation

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \to \mathcal{N}(0,1) \quad \text{(in distribution as } n \to \infty).$$

These results provide the foundation for point estimation, confidence intervals, and hypothesis testing developed in the following chapters.

Chapter 2

Point Estimation

2.1 Introduction to Point Estimation

In statistical inference, one of the main objectives is to learn about unknown characteristics of a population from a sample. When the population distribution is described by a model involving unknown parameters, these parameters must be estimated from the available observations.

Point estimation is concerned with assigning a single numerical value to an unknown parameter on the basis of a random sample.

Let

$$X_1, \ldots, X_n$$

be a random sample from a population whose distribution depends on an unknown parameter θ. A **point estimator** is a statistic constructed from the sample and used to estimate θ.

Point estimation has two complementary aspects. The first is the construction of estimators, while the second is the study of their properties. Two important methods for constructing estimators are the Maximum Likelihood Method and the Method of Moments. Once an estimator has been obtained, its bias, variance, efficiency, consistency, sufficiency, completeness, and other properties can be investigated.

Point estimation should be distinguished from confidence interval estimation. A point estimator provides a single value, whereas a confidence interval provides a range of values associated with a specified confidence level. Confidence interval estimation will be studied in Chapter 3.

2.2 Parametric and Nonparametric Estimation

Statistical estimation may be considered within either a parametric or a nonparametric framework.

2.2.1 Parametric Estimation

In parametric estimation, the probability distribution of the population is assumed to belong to a specified family of distributions characterized by a finite-dimensional parameter.

A parametric statistical model can be represented as

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where Θ is the parameter space and θ denotes the unknown parameter.

For example, suppose that

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

If σ^2 is known and μ is unknown, the parameter to be estimated is $\theta = \mu$. If both μ and σ^2 are unknown, the parameter is the vector

$$\theta = (\mu, \sigma^2).$$

The advantage of the parametric approach is that the assumed model provides a mathematical structure that can be used to construct and study estimators.

2.2.2 Nonparametric Estimation

In nonparametric estimation, the distribution is not restricted to a finite-dimensional parametric family of the type considered above. The object of estimation may be, for example, a distribution function, a density function, or another characteristic of the population without specifying a particular finite-dimensional distributional model.

Thus, the main distinction is that parametric estimation is based on a finite-dimensional parameterization of the statistical model, whereas nonparametric estimation does not impose such a restriction. The present chapter focuses mainly on point estimation in parametric models.

2.3 Point Estimators and Estimates

Let

$$X_1, \ldots, X_n$$

be a random sample and let θ be an unknown parameter.

A **point estimator** of θ is a statistic

$$T = T(X_1, \ldots, X_n)$$

constructed from the random sample with the purpose of estimating θ. Since T is a function of random variables, it is itself a random variable. Its value is therefore not known before the sample is observed.

After the observations

$$x_1, \ldots, x_n$$

have been obtained, the estimator takes the numerical value

$$t = T(x_1, \ldots, x_n).$$

This numerical value is called the **estimate** of θ.

For example, suppose that μ is the unknown mean of a population. The sample mean

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

is an estimator of μ. Once the observations $x_1, \ldots, x_n$ are available, the corresponding estimate is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Therefore, the estimator and the estimate are different objects. The former is random because it depends on the random sample, whereas the latter is a fixed numerical value after the observations have been collected.

2.4 Random Variables, Estimators, and Unknown Parameters

A clear distinction between random variables, estimators, and parameters is essential in statistical inference.

The observations $X_1, \ldots, X_n$ are random variables before the experiment is performed. Their observed values $x_1, \ldots, x_n$ are fixed numbers. An estimator

$$T(X_1, \ldots, X_n)$$

is also a random variable because it is calculated from the random sample.

In the classical frequentist framework, the parameter θ is regarded as a fixed but unknown quantity. It is not a random variable. Consequently, before observing the sample,

$$X_1, \ldots, X_n \quad \text{and} \quad T(X_1, \ldots, X_n)$$

are random, while θ is fixed but unknown. After the observations have been collected, the sample values and the resulting estimate are fixed numerical quantities.

This distinction becomes particularly important when defining the bias, variance, and other properties of an estimator.

2.5 Unknown Parameters

Consider a statistical model with density or probability mass function

$$f(x; \theta), \qquad \theta \in \Theta.$$

The parameter θ describes a characteristic of the population distribution. Depending on the model, it may represent a mean, a variance, a probability, a rate, or several parameters simultaneously.

Suppose that

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x; \theta).$$

The parameter θ is unknown, while the observations are available. Point estimation seeks to construct a statistic

$$T(X_1, \ldots, X_n)$$

that provides useful information about θ.

The quality of an estimator cannot generally be determined from one observed sample alone. Its behavior under repeated sampling must be considered. This leads to the study of the properties of estimators.

2.6 Methods of Point Estimation

Two fundamental methods of point estimation considered in this course are:

1. the Maximum Likelihood Method;
2. the Method of Moments.

The Maximum Likelihood Method constructs an estimator by maximizing the likelihood of the observed data as a function of the unknown parameter. The Method of Moments constructs estimators by equating sample moments to the corresponding theoretical moments.

2.6.1 An Introductory Example

Suppose that

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p),$$

where p is unknown. For each observation,

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \qquad x_i \in \{0, 1\}.$$

A natural estimator of p is the proportion of observations equal to one:

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

This example illustrates the fundamental idea of point estimation: an unknown population parameter is replaced by a statistic calculated from the sample. The same estimator can also be obtained systematically by both the Maximum Likelihood Method and the Method of Moments.

2.6.2 Maximum Likelihood Method

Let

$$X_1, \ldots, X_n$$

be independent observations from a model with density or probability mass function

$$f(x; \theta), \qquad \theta \in \Theta.$$

After observing $x_1, \ldots, x_n$, the likelihood function is defined by

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta),$$

when the observations are independent. The likelihood is regarded as a function of θ, with the observed data kept fixed.

A maximum likelihood estimator is a value of θ that maximizes the likelihood over the parameter space:

$$\widehat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n).$$

The maximum may occur at an interior point of the parameter space or at its boundary. Therefore, solving a likelihood equation is not by itself sufficient; the resulting candidate must be checked against the parameter space and the behavior of the likelihood.

2.6.3 The Log-Likelihood Function

Because products can be difficult to manipulate, it is often more convenient to work with the logarithm of the likelihood. The log-likelihood function is

$$\ell(\theta) = \log L(\theta; x_1, \ldots, x_n).$$

Since the logarithm is strictly increasing,

$$\arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Thus, maximizing the log-likelihood gives the same maximum likelihood estimator while usually simplifying differentiation and algebra. For an interior maximum, a candidate estimator can often be obtained by solving

$$\frac{\partial}{\partial \theta} \ell(\theta) = 0.$$

The second derivative or another appropriate argument is then used to verify that the candidate corresponds to a maximum.

2.6.4 Likelihood Functions for Common Models

The form of the likelihood depends on the statistical model. For a Bernoulli sample,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p),$$

the likelihood is

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}.$$

Therefore,

$$L(p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}.$$

The log-likelihood is

$$\ell(p) = \left(\sum_{i=1}^{n} x_i \right) \log p + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-p).$$

Differentiating gives

$$\ell'(p) = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}.$$

Solving $\ell'(p) = 0$ gives

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

Hence the maximum likelihood estimator is

$$\widehat{p}_{\text{MLE}} = \overline{X}.$$

Other common models, such as the normal, exponential, Poisson, and uniform models, lead to likelihood functions of different forms. The general principle remains the same: construct the likelihood from the statistical model and maximize it with respect to the unknown parameter.

2.6.5 Maximum Likelihood Estimator

The maximum likelihood estimator is therefore obtained by

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} L(\theta).$$

When the likelihood has a unique maximum, the MLE is uniquely determined. Maximum likelihood estimation is widely used because it provides a general method that can be applied to a large class of statistical models. However, maximum likelihood estimators are not automatically unbiased, efficient, or consistent in every problem. These properties must be studied for the particular model under consideration.

2.6.6 Method of Moments

The Method of Moments is another general procedure for constructing estimators. Let the theoretical moments of a random variable X be

$$\mu_r'(\theta) = \mathbb{E}_\theta[X^r], \qquad r = 1, 2, \ldots$$

The corresponding sample moments are

$$m_r' = \frac{1}{n} \sum_{i=1}^{n} X_i^r.$$

The Method of Moments consists in equating theoretical moments to sample moments. For a model containing one unknown parameter, the first moment equation is typically

$$\mu_1'(\theta) = m_1'.$$

Solving this equation gives the Method of Moments estimator.

If the parameter is multidimensional, several moment equations can be used. For a parameter vector

$$\theta = (\theta_1, \ldots, \theta_k),$$

one may use the first k moment equations:

$$\mu_r'(\theta) = m_r', \qquad r = 1, \ldots, k.$$

The resulting system is solved for the unknown parameters. The Method of Moments and the Maximum Likelihood Method may lead to the same estimator in some models, but they are different estimation procedures and need not produce identical results.

2.7 Properties of Estimators

After constructing an estimator, the next question is how well it performs. Different estimators of the same parameter can differ in their bias, variability, efficiency, and asymptotic behavior. The properties introduced in this section provide criteria for studying and comparing estimators.

2.7.1 Bias

Let T be an estimator of θ. Its bias at θ is defined by

$$\text{Bias}_\theta(T) = \mathbb{E}_\theta[T] - \theta,$$

provided the expectation exists.

An estimator is called **unbiased** if

$$\mathbb{E}_\theta[T] = \theta$$

for every $\theta \in \Theta$.

For example, if $X_1, \ldots, X_n$ have mean μ, then

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

Therefore, $\overline{X}$ is an unbiased estimator of μ.

Bias measures systematic deviation of the average value of an estimator from the parameter being estimated.

2.7.2 Asymptotic Bias

Let T_n be an estimator of θ based on a sample of size n. Its bias is

$$\text{Bias}_\theta(T_n) = \mathbb{E}_\theta[T_n] - \theta.$$

If

$$\lim_{n\to\infty} \text{Bias}_\theta(T_n) = 0,$$

then the bias vanishes asymptotically.

An estimator may therefore be biased for finite samples while becoming asymptotically unbiased. Unbiasedness is a finite-sample property, whereas asymptotic unbiasedness concerns the limiting behavior of the estimator.

2.7.3 Variance

The variance of an estimator T is

$$\text{Var}_\theta(T) = \mathbb{E}_\theta\left[(T - \mathbb{E}_\theta[T])^2\right].$$

Variance measures the variability of the estimator under repeated sampling. For independent observations with variance σ^2,

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Thus, increasing the sample size generally reduces the variability of the sample mean under the usual independent sampling assumptions.

Bias and variance measure different aspects of estimator performance. An estimator can have a small variance but a nonzero bias, or it can be unbiased while having a relatively large variance.

2.7.4 Mean Squared Error

The mean squared error of an estimator T of θ is

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta\left[(T-\theta)^2\right].$$

It can be decomposed as

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + \mathrm{Bias}_\theta(T)^2.$$

This decomposition shows that estimator performance can involve a trade-off between variability and systematic error.

2.7.5 Efficiency

Efficiency provides a way of comparing estimators of the same parameter. Suppose that T_1 and T_2 are unbiased estimators of θ. If

$$\mathrm{Var}_\theta(T_1) \le \mathrm{Var}_\theta(T_2)$$

for all $\theta \in \Theta$, then T_1 has no greater variance than T_2.

When the inequality is strict for some parameter values, the two estimators do not have the same efficiency over the whole parameter space. Efficiency will later be related to Fisher information and the Cramér–Rao–Fréchet–Darmois bound.

2.7.6 Admissibility

Admissibility is a decision-theoretic property. Suppose that an estimator is evaluated through a loss function and the corresponding risk function. An estimator T_1 is said to dominate an estimator T_2 if its risk is no larger for every parameter value and strictly smaller for at least one parameter value.

An estimator is called **admissible** if no other estimator dominates it under the specified decision criterion. Admissibility is therefore different from unbiasedness and consistency. It depends on the loss function used to evaluate the statistical procedure.

2.7.7 Consistency

Let T_n be an estimator of θ based on a sample of size n. The estimator sequence (T_n) is said to be **consistent** for θ if

$$T_n \xrightarrow{\mathbb{P}} \theta \qquad \text{as } n \to \infty.$$

Consistency means that the estimator becomes increasingly concentrated around the true parameter as the sample size grows. An estimator does not need to be unbiased for every finite sample size in order to be consistent.

2.7.8 Convergence in Probability

The notation

$$T_n \xrightarrow{\mathbb{P}} \theta$$

means that, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P_\theta\left(|T_n - \theta| > \varepsilon\right) = 0.$$

This is called **convergence in probability**. Convergence in probability is the principal mode of convergence used in the definition of consistency.

For example, the law of large numbers implies, under appropriate conditions, that

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

Thus, the sample mean is consistent for the population mean under those conditions.

2.7.9 Convergence in Mean of Order k

Let $k > 0$. A sequence of random variables (T_n) is said to converge in mean of order k to θ if

$$\lim_{n\to\infty} \mathbb{E}_\theta\left[|T_n - \theta|^k\right] = 0.$$

This is commonly denoted by

$$T_n \xrightarrow{L^k} \theta.$$

For $k = 2$, the convergence is called convergence in quadratic mean. Under the usual conditions, convergence in mean of order k implies convergence in probability. Thus, convergence in mean provides a stronger type of control over the distance between the estimator and the parameter.

2.8 Sufficiency

A statistic can summarize the sample while retaining all the information about a parameter that is relevant for statistical inference.

2.8.1 Formal Definition

Let

$$\mathbf{X} = (X_1, \ldots, X_n)$$

have a distribution depending on θ, and let

$$T = T(\mathbf{X})$$

be a statistic.

Definition 2.1. The statistic T is called **sufficient** (or **exhaustive**) for θ if the conditional distribution of the sample given T does not depend on θ. Formally:

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n \mid T = t) \quad \text{is independent of } \theta.$$

Interpretation: Once T is observed, the sample provides no further information about θ. T captures all the information about θ contained in the data.

For models having a joint density or probability mass function, sufficiency can often be established using the factorization criterion.

2.8.2 Fisher–Neyman Factorization Theorem

Theorem 2.1 (Fisher–Neyman Factorization Theorem)**.** *A statistic $T(X_1, \ldots, X_n)$ is sufficient for θ if and only if the joint density or probability mass function can be factored as:*

$$f(x_1, \ldots, x_n; \theta) = g\big(T(x_1, \ldots, x_n), \theta\big)\, h(x_1, \ldots, x_n),$$

where:

- $g(t, \theta) \geq 0$ *depends on the data only through* T*;*
- $h(x_1, \ldots, x_n) \geq 0$ *does not depend on* θ*.*

This theorem provides the most practical tool to establish sufficiency.

2.8.3 Proof of the Factorization Theorem

We prove this in the discrete case. Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote the random vector and $\mathbf{x} = (x_1, \ldots, x_n)$ a realization.

Direct Implication: Assume the factorization $f(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)\, h(\mathbf{x})$ holds. We show that T is sufficient.

Step 1: Express the event $\{T = t\}$ in terms of $\mathbf{X}$:

$$\{T(\mathbf{X}) = t\} = \bigcup_{\mathbf{x}' : T(\mathbf{x}') = t} \{\mathbf{X} = \mathbf{x}'\}.$$

Step 2: The union is disjoint because for $\mathbf{x}'_1 \neq \mathbf{x}'_2$, the events $\{\mathbf{X} = \mathbf{x}'_1\}$ and $\{\mathbf{X} = \mathbf{x}'_2\}$ are mutually exclusive.

Step 3: By σ-additivity of probability:

$$P_\theta(T = t) = \sum_{\mathbf{x}' : T(\mathbf{x}') = t} P_\theta(\mathbf{X} = \mathbf{x}').$$

Step 4: Using the joint pmf $f(\mathbf{x}'; \theta) = P_\theta(\mathbf{X} = \mathbf{x}')$ and independence:

$$P_\theta(T = t) = \sum_{\mathbf{x}' : T(\mathbf{x}') = t} \prod_{i=1}^{n} f(x'_i; \theta).$$

Step 5: For any $\mathbf{x}$ such that $T(\mathbf{x}) = t$:

$$P_\theta(\mathbf{X} = \mathbf{x} \mid T = t) = \frac{P_\theta(\mathbf{X} = \mathbf{x})}{P_\theta(T = t)} = \frac{g(t, \theta)\, h(\mathbf{x})}{\displaystyle\sum_{\mathbf{x}' : T(\mathbf{x}') = t} g(t, \theta)\, h(\mathbf{x}')}.$$

The factor $g(t, \theta)$ cancels:

$$P_\theta(\mathbf{X} = \mathbf{x} \mid T = t) = \frac{h(\mathbf{x})}{\displaystyle\sum_{\mathbf{x}' : T(\mathbf{x}') = t} h(\mathbf{x}')}.$$

This expression is independent of θ. Therefore, T is sufficient.

Converse Implication: Assume T is sufficient.

Step 1: By definition,

$$P_\theta(\mathbf{X} = \mathbf{x} \mid T = t) = h(\mathbf{x}),$$

where h is independent of θ.

Step 2: Using conditional probability:

$$f(\mathbf{x}; \theta) = P_\theta(\mathbf{X} = \mathbf{x} \mid T = T(\mathbf{x}))\, P_\theta(T = T(\mathbf{x})).$$

Step 3: Define $g(t, \theta) := P_\theta(T = t)$. Then:

$$f(\mathbf{x}; \theta) = g\big(T(\mathbf{x}), \theta\big)\, h(\mathbf{x}). \quad \blacksquare$$

2.8.4 Examples of Sufficient Statistics

Example 2.1 (Bernoulli Model)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. The joint pmf is:

$$f(x_1, \ldots, x_n; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

Let $T = \sum_{i=1}^n X_i$. We can factor:

$$f(\mathbf{x}; p) = \underbrace{p^T(1-p)^{n-T}}_{g(T,p)} \cdot \underbrace{1}_{h(\mathbf{x})}.$$

By the factorization theorem, $T = \sum_{i=1}^n X_i$ is sufficient for p.

Example 2.2 (Poisson Model)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. The joint pmf is:

$$f(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda}\, \lambda^{\sum_{i=1}^n x_i} \prod_{i=1}^{n} \frac{1}{x_i!}.$$

Let $T = \sum_{i=1}^n X_i$. We can factor:

$$f(\mathbf{x}; \lambda) = \underbrace{e^{-n\lambda}\, \lambda^T}_{g(T,\lambda)} \cdot \underbrace{\prod_{i=1}^{n} \frac{1}{x_i!}}_{h(\mathbf{x})}.$$

By the factorization theorem, $T = \sum_{i=1}^n X_i$ is sufficient for λ.

Example 2.3 (Exponential Model)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$. The joint density is:

$$f(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \lambda\, e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \cdot \mathbf{1}_{\{x_i > 0,\, \forall i\}}.$$

Let $T = \sum_{i=1}^n X_i$. We can factor:

$$f(\mathbf{x}; \lambda) = \underbrace{\lambda^n\, e^{-\lambda T}}_{g(T,\lambda)} \cdot \underbrace{\mathbf{1}_{\{x_i > 0,\, \forall i\}}}_{h(\mathbf{x})}.$$

By the factorization theorem, $T = \sum_{i=1}^n X_i$ is sufficient for λ.

Example 2.4 (Normal Model)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. The joint density is:

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

This can be rewritten as:

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

Expanding the sum:

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2.$$

Therefore, the joint density factors as:

$$f(\mathbf{x}; \mu, \sigma^2) = \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2\right)}_{g(T(\mathbf{x}), \mu, \sigma^2)} \cdot \underbrace{1}_{h(\mathbf{x})},$$

where $T(\mathbf{x}) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$.

By the factorization theorem, the statistic

$$T = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$$

is sufficient for (μ, σ^2).

2.9 Completeness

2.9.1 Formal Definition

Let T be a statistic for a family of distributions $\{P_\theta : \theta \in \Theta\}$.

Definition 2.2. T is said to be **complete** if for every measurable function g:

$$\mathbb{E}_\theta[g(T)] = 0 \quad \text{for all } \theta \in \Theta \quad \Longrightarrow \quad P_\theta(g(T) = 0) = 1 \quad \text{for all } \theta \in \Theta.$$

Interpretation: The only function g such that $\mathbb{E}_\theta[g(T)] = 0$ for all θ is the zero function, up to P_θ-null sets. Completeness ensures uniqueness of unbiased estimators based on T.

If T is both sufficient and complete, then any unbiased estimator based on T is unique. This means: if $\hat{\theta}_1(T)$ and $\hat{\theta}_2(T)$ are both unbiased for θ, then

$$\hat{\theta}_1(T) = \hat{\theta}_2(T) \quad \text{almost surely.}$$

Completeness prevents the existence of "different" unbiased estimators based on the same sufficient statistic. This property is crucial for the Lehmann–Scheffé theorem.

2.9.2 Proofs of Completeness for Common Models

Lemma 2.1. *Let* $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$*, with* $p \in (0,1)$*. Then* $T = \sum_{i=1}^n X_i$ *is complete for* p*.*

Proof. **Step 1:** $T \sim \text{Binomial}(n, p)$, with pmf

$$P_p(T = t) = \binom{n}{t} p^t (1-p)^{n-t}, \qquad t \in \{0, 1, \ldots, n\}.$$

Step 2: Suppose $\mathbb{E}_p[g(T)] = 0$ for all $p \in (0,1)$. Then:

$$\sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} = 0, \qquad \forall\, p \in (0,1).$$

Step 3: Multiply both sides by $\dfrac{1}{(1-p)^n}$ (valid since $p \in (0,1)$):

$$\sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = 0, \qquad \forall\, p \in (0,1).$$

Step 4: Use the substitution $u = \dfrac{p}{1-p}$. As p ranges over $(0,1)$, u ranges over $(0, +\infty)$. We obtain a polynomial in u:

$$\sum_{t=0}^{n} g(t) \binom{n}{t} u^t = 0, \qquad \forall\, u > 0.$$

Step 5: A polynomial that vanishes on an infinite set must be identically zero. Therefore, all its coefficients are zero:

$$g(t) \binom{n}{t} = 0 \quad \text{for all } t \in \{0, 1, \ldots, n\}.$$

Since $\binom{n}{t} > 0$, we conclude:

$$g(t) = 0 \quad \text{for all } t. \quad \blacksquare$$

□

Lemma 2.2. *Let* $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$*, with* $\lambda > 0$*. Then* $T = \sum_{i=1}^n X_i$ *is complete for* λ*.*

Proof. **Step 1:** $T \sim \text{Poisson}(n\lambda)$, with pmf

$$P_\lambda(T = t) = \frac{e^{-n\lambda}(n\lambda)^t}{t!}, \qquad t \in \{0, 1, 2, \ldots\}.$$

Step 2: Suppose $\mathbb{E}_\lambda[g(T)] = 0$ for all $\lambda > 0$. Then:

$$\sum_{t=0}^{\infty} g(t) \frac{e^{-n\lambda}(n\lambda)^t}{t!} = 0, \qquad \forall\, \lambda > 0.$$

Step 3: Multiply both sides by $e^{n\lambda}$:

$$\sum_{t=0}^{\infty} g(t)\frac{(n\lambda)^t}{t!} = 0, \qquad \forall\, \lambda > 0.$$

This is a power series in λ that vanishes for all $\lambda > 0$.

Step 4: Use the identity theorem for power series. A power series that vanishes on an interval $(0, +\infty)$ must be identically zero. Therefore, all its coefficients are zero:

$$g(t)\frac{n^t}{t!} = 0 \quad \text{for all } t \in \{0, 1, 2, \ldots\}.$$

Step 5: Since $\dfrac{n^t}{t!} > 0$ for all t, we conclude:

$$g(t) = 0 \quad \text{for all } t \in \{0, 1, 2, \ldots\}. \quad \blacksquare$$

□

Lemma 2.3. *Let* $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$*, with* $\lambda > 0$*. Then* $T = \sum_{i=1}^{n} X_i$ *is complete for* λ*.*

Proof. **Step 1:** $T \sim \text{Gamma}(n, \lambda)$, with density

$$f_T(t; \lambda) = \frac{\lambda^n}{(n-1)!}\, t^{n-1}\, e^{-\lambda t}, \qquad t > 0.$$

Step 2: Suppose $\mathbb{E}_\lambda[g(T)] = 0$ for all $\lambda > 0$. Then:

$$\int_0^{\infty} g(t)\frac{\lambda^n}{(n-1)!}\, t^{n-1}\, e^{-\lambda t}\, dt = 0, \qquad \forall\, \lambda > 0.$$

Step 3: Factor out the constants (independent of t) and multiply both sides by $\dfrac{(n-1)!}{\lambda^n}$:

$$\int_0^{\infty} g(t)\, t^{n-1}\, e^{-\lambda t}\, dt = 0, \qquad \forall\, \lambda > 0.$$

Step 4: Recognize the Laplace transform. The left-hand side is exactly the Laplace transform of the function $h(t) = g(t)\, t^{n-1}$ evaluated at λ:

$$\mathcal{L}[h](\lambda) = \int_0^{\infty} h(t)\, e^{-\lambda t}\, dt = 0, \qquad \forall\, \lambda > 0.$$

Step 5: Use the uniqueness theorem for Laplace transforms: if $\mathcal{L}[h](\lambda) = 0$ for all $\lambda > 0$, then $h(t) = 0$ almost everywhere.

Therefore, $g(t)\, t^{n-1} = 0$ a.e. Since $t^{n-1} > 0$ for $t > 0$, we conclude:

$$g(t) = 0 \quad \text{a.e. on } (0, +\infty). \quad \blacksquare$$

□

2.9.3 Completeness and Exponential Families

To establish the completeness of statistics for many common models, it is highly useful to rely on the properties of exponential families.

Theorem 2.2 (Completeness in Exponential Families)**.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample from a k-parameter exponential family with natural parameters $(\eta_1, \ldots, \eta_k)$ and natural sufficient statistics $(T_1, \ldots, T_k)$.*

If the natural parameter space

$$\mathcal{H} = \left\{ (\eta_1, \ldots, \eta_k) : \int \exp\left(\sum_{j=1}^{k} \eta_j T_j(x) \right) h(x)\, dx < \infty \right\}$$

*contains an open rectangle in $\mathbb{R}^k$ (i.e., the family is of **full rank** or **regular**), then the statistic*

$$T = (T_1, \ldots, T_k)$$

*is **complete** for $(\eta_1, \ldots, \eta_k)$.*

This powerful theorem allows us to establish completeness without resorting to direct and often difficult analytical calculations (such as Laplace transforms or power series), provided the model belongs to a regular exponential family.

Lemma 2.4. *Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. Then the statistic*

$$T = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right)$$

is complete for (μ, σ^2).

Proof. **Step 1:** The normal model belongs to the exponential family. The joint density can be written in canonical form:

$$f(\mathbf{x}; \mu, \sigma^2) = \exp\left(\frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - \frac{n\mu^2}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2) \right).$$

The natural parameters are:

$$\eta_1 = \frac{\mu}{\sigma^2} \quad \text{and} \quad \eta_2 = -\frac{1}{2\sigma^2},$$

and the corresponding sufficient statistics are:

$$T_1 = \sum_{i=1}^{n} X_i \quad \text{and} \quad T_2 = \sum_{i=1}^{n} X_i^2.$$

Step 2: The parameter space for (η_1, η_2) is:

$$\left\{ (\eta_1, \eta_2) \in \mathbb{R}^2 : \eta_2 < 0 \right\},$$

which contains an open rectangle in $\mathbb{R}^2$.

Step 3: By the general theorem on exponential families (stated in the previous subsection), when the natural parameter space contains an open set in $\mathbb{R}^k$ (here $k = 2$), the family is of full rank, and the natural sufficient statistic $T = (T_1, T_2)$ is **complete**.

Therefore, $T = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is complete for (μ, σ^2). ■ □

2.10 Usual Estimators

Several estimators occur repeatedly in statistical applications. The most common examples concern quantitative variables and qualitative variables.

2.10.1 Quantitative Case

Let

$$X_1, \ldots, X_n$$

be observations of a quantitative variable.

The sample mean is

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

It is the usual estimator of the population mean μ.

The usual unbiased estimator of the population variance σ^2 is

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

The corresponding sample standard deviation is

$$S = \sqrt{S^2}.$$

The denominator $n-1$ is important in the unbiased estimation of σ^2 under the standard independent sampling assumptions.

2.10.2 Qualitative Case: Unknown Proportion

Suppose that each observation belongs to one of two categories. Define

$$X_i = \begin{cases} 1, & \text{if the observation belongs to the category of interest,} \\ 0, & \text{otherwise.} \end{cases}$$

If

$$X_i \sim \text{Bernoulli}(p),$$

then p represents the population proportion of observations belonging to the category of interest.

The natural estimator of p is the sample proportion

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

In the Bernoulli model, this estimator is also the maximum likelihood estimator of p.

2.11 Fisher Information and Score

The likelihood contains information about the unknown parameter. The score function and Fisher information provide two important tools for quantifying this information.

2.11.1 The Score Function

Let X have density or probability mass function

$$f(x;\theta).$$

The **score function** is defined by

$$U(\theta;X) = \frac{\partial}{\partial\theta}\log f(X;\theta),$$

whenever the derivative exists.

For a sample of independent observations $X_1,\dots,X_n$, the sample score is

$$U_n(\theta) = \frac{\partial}{\partial\theta}\log L(\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta).$$

The score describes the local sensitivity of the log-likelihood to changes in the parameter.

Key Property: Under regularity conditions, the score has zero expectation:

$$\mathbb{E}_\theta[U(\theta;X)] = 0, \qquad \forall\,\theta\in\Theta.$$

Proof sketch: Since $\int f(x;\theta)\,dx = 1$, differentiating both sides with respect to θ:

$$\int \frac{\partial}{\partial\theta} f(x;\theta)\,dx = 0.$$

Rewriting using $\dfrac{\partial}{\partial\theta}\log f = \dfrac{1}{f}\dfrac{\partial f}{\partial\theta}$:

$$\int \frac{\partial}{\partial\theta}\log f(x;\theta)\cdot f(x;\theta)\,dx = \mathbb{E}_\theta[U(\theta;X)] = 0. \quad \blacksquare$$

Consequence: The MLE satisfies $U_n(\widehat{\theta}_{\text{MLE}}) = 0$ (first-order condition).

2.11.2 Fisher Information

Under appropriate regularity conditions, the Fisher information in one observation is

$$I(\theta) = \mathbb{E}_\theta\left[U(\theta;X)^2\right].$$

Under the same type of regularity conditions, it can also be written as

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

For independent observations,

$$I_n(\theta) = n\,I(\theta).$$

Thus, the total information increases linearly with the sample size in the independent case.

Interpretation: Fisher information measures the amount of information about θ contained in a single observation. Higher information $\Rightarrow$ more precise estimation.

2.11.3 Example: Fisher Information for Exponential Model

Let $X \sim \text{Exp}(\lambda)$. The density is $f(x;\lambda) = \lambda\, e^{-\lambda x}$ for $x > 0$.

Step 1: Log-density:

$$\log f(x;\lambda) = \log\lambda - \lambda x.$$

Step 2: Score function:

$$U(\lambda; X) = \frac{\partial}{\partial\lambda} \log f(X;\lambda) = \frac{1}{\lambda} - X.$$

Step 3: Fisher information:

$$I(\lambda) = \text{Var}_\lambda\left(\frac{1}{\lambda} - X\right) = \text{Var}_\lambda(X) = \frac{1}{\lambda^2}.$$

For a sample of size n:

$$I_n(\lambda) = \frac{n}{\lambda^2}.$$

2.11.4 Efficiency and Fisher Information

Fisher information provides a theoretical benchmark for the precision of estimators. For a regular one-parameter model, the Cramér–Rao–Fréchet–Darmois bound uses Fisher information to provide a lower bound for the variance of an unbiased estimator.

This establishes an important connection:

$$\text{Fisher information} \longrightarrow \text{variance bound} \longrightarrow \text{efficiency}.$$

An estimator that attains the relevant lower bound is efficient in the Cramér–Rao sense.

2.11.5 Fisher Information and Completeness

Fisher information and completeness are distinct concepts. Fisher information is a measure of the local information about a parameter contained in a statistical model. Completeness is a property of the family of distributions of a statistic. It is mainly connected with uniqueness properties of unbiased estimators.

Both concepts are important in the study of estimators, but they should not be interpreted as equivalent notions.

2.12 Cramér–Rao–Fréchet–Darmois Bound

The Cramér–Rao–Fréchet–Darmois bound provides a lower bound for the variance of unbiased estimators under suitable regularity conditions.

Consider a regular one-parameter model with Fisher information $I_n(\theta)$ based on a sample of size n. If T is an unbiased estimator of θ, then, under the required regularity conditions,

$$\text{Var}_\theta(T) \geq \frac{1}{I_n(\theta)}.$$

For independent observations,

$$I_n(\theta) = n\, I(\theta),$$

so that

$$\mathrm{Var}_\theta(T) \geq \frac{1}{n\, I(\theta)}.$$

The bound provides a benchmark against which the variance of an unbiased estimator can be compared. If an unbiased estimator attains the bound, then it has the smallest variance allowed by the bound under the stated conditions and is efficient in the Cramér–Rao sense.

The regularity assumptions are essential. The bound cannot be applied mechanically to every statistical model.

2.12.1 Illustrative Example: Normal Mean

Suppose that

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

where σ^2 is known and μ is unknown.

For one observation,

$$\log f(X; \mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}.$$

The score function is

$$U(\mu; X) = \frac{X-\mu}{\sigma^2}.$$

Consequently,

$$I(\mu) = \mathbb{E}_\mu\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}.$$

For n independent observations,

$$I_n(\mu) = \frac{n}{\sigma^2}.$$

The Cramér–Rao bound therefore gives

$$\mathrm{Var}_\mu(T) \geq \frac{\sigma^2}{n}$$

for every unbiased estimator T of μ satisfying the required conditions.

On the other hand,

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

Thus, the sample mean attains the bound in this model and is efficient in the Cramér–Rao sense.

2.12.2 Minimum Variance Unbiased Estimator (MVUE)

The official program mentions the concept of "ESBVM" (in French: *Estimateur Sans Biais de Variance Minimale*), which translates to **Minimum Variance Unbiased Estimator (MVUE)** in English.

An unbiased estimator T of θ is called an MVUE if, for any other unbiased estimator T' of θ, we have

$$\mathrm{Var}_\theta(T) \leq \mathrm{Var}_\theta(T')$$

for all $\theta \in \Theta$.

The search for an MVUE is greatly facilitated by combining the concepts of sufficiency and completeness. The fundamental result in this regard is the **Lehmann–Scheffé Theorem**:

Theorem 2.3 (Lehmann–Scheffé)**.** *If T is a complete and sufficient statistic for θ, and if $g(T)$ is an unbiased estimator of θ, then $g(T)$ is the unique MVUE of θ.*

Proof. The proof combines two fundamental ingredients: Sufficiency (Rao–Blackwell theorem) and Completeness (Uniqueness).

Setup: Let $\tilde{\theta}$ be any unbiased estimator of θ. We define:

$$\hat{\theta}(T) := \mathbb{E}_\theta[\tilde{\theta} \mid T].$$

Since T is sufficient, $\hat{\theta}(T)$ does not depend on θ.

Step 1 (Rao–Blackwell): $\hat{\theta}(T)$ is unbiased and depends only on T.
By the tower property of conditional expectation:

$$\mathbb{E}_\theta\big[\hat{\theta}(T)\big] = \mathbb{E}_\theta\big[\mathbb{E}_\theta[\tilde{\theta} \mid T]\big] = \mathbb{E}_\theta[\tilde{\theta}] = \theta.$$

So $\hat{\theta}(T)$ is an unbiased estimator of θ. Since T is sufficient, the conditional distribution of the sample given T does not depend on θ. Therefore:

$$\hat{\theta}(T) = \mathbb{E}_\theta[\tilde{\theta} \mid T]$$

is a function of T alone, independent of θ.

Step 2 (Variance Reduction): The variance of $\hat{\theta}(T)$ is smaller than or equal to that of $\tilde{\theta}$.

Law of total variance (general formula): For any two random variables Y and X:

$$\mathrm{Var}(Y) = \mathbb{E}\big[\mathrm{Var}(Y \mid X)\big] + \mathrm{Var}\big(\mathbb{E}[Y \mid X]\big).$$

Application to our case (with $Y = \tilde{\theta}$ and $X = T$):

$$\mathrm{Var}_\theta(\tilde{\theta}) = \underbrace{\mathrm{Var}_\theta\big(\mathbb{E}_\theta[\tilde{\theta} \mid T]\big)}_{=\mathrm{Var}_\theta(\hat{\theta}(T))} + \underbrace{\mathbb{E}_\theta\big[\mathrm{Var}_\theta(\tilde{\theta} \mid T)\big]}_{\geq 0}.$$

Since the second term is non-negative:

$$\mathrm{Var}_\theta(\tilde{\theta}) \geq \mathrm{Var}_\theta\big(\hat{\theta}(T)\big), \qquad \forall\, \theta \in \Theta.$$

Conclusion of Step 2: $\hat{\theta}(T)$ has smallest variance among all unbiased estimators.

Step 3 (Uniqueness via Completeness): Suppose $\hat{\theta}_1(T)$ and $\hat{\theta}_2(T)$ are two unbiased estimators of θ based on T, both attaining the minimum variance.

Define:

$$\Delta(T) := \hat{\theta}_1(T) - \hat{\theta}_2(T).$$

Since both are unbiased:

$$\mathbb{E}_\theta[\Delta(T)] = \mathbb{E}_\theta[\hat{\theta}_1(T)] - \mathbb{E}_\theta[\hat{\theta}_2(T)] = \theta - \theta = 0, \qquad \forall\, \theta \in \Theta.$$

Since T is complete, this implies:

$$P_\theta(\Delta(T) = 0) = 1, \qquad \forall\, \theta \in \Theta.$$

Therefore, $\hat{\theta}_1(T) = \hat{\theta}_2(T)$ almost surely. ■ □

This theorem establishes a powerful link: once a complete sufficient statistic is found, any unbiased estimator based solely on it is guaranteed to be optimal in terms of variance among all unbiased estimators.

2.13 Conclusion

Point estimation provides a systematic framework for learning about unknown population parameters from a sample.

The chapter has introduced the distinction between parametric and nonparametric estimation, point estimators and estimates, and random variables and unknown parameters. Two important methods for constructing estimators were then presented: Maximum Likelihood and the Method of Moments.

The statistical properties of estimators were subsequently introduced, including bias, asymptotic bias, variance, efficiency, admissibility, consistency, sufficiency, completeness, convergence in probability, and convergence in mean of order k.

The usual estimators of a population mean, variance, standard deviation, and proportion were also considered.

Finally, the score function and Fisher information were introduced and linked to the Cramér–Rao–Fréchet–Darmois bound, estimator efficiency, and the concept of the Minimum Variance Unbiased Estimator (MVUE) via the Lehmann–Scheffé theorem.

These concepts form the theoretical foundation for confidence interval estimation and statistical testing, which will be developed in the following chapters.

Chapter 3

Confidence Interval Estimation

3.1 Point Estimation versus Confidence Interval Estimation

In point estimation, an unknown population parameter is estimated by a single numerical value obtained from a sample.

Let

$$X_1, \ldots, X_n$$

be a random sample from a population depending on an unknown parameter θ. A point estimator of θ is a statistic

$$T = T(X_1, \ldots, X_n).$$

After observing the sample, the estimator takes a numerical value

$$t = T(x_1, \ldots, x_n),$$

which is used as an estimate of θ.

Limitation of Point Estimation: A point estimate does not, however, indicate the uncertainty associated with the estimation. Two samples drawn from the same population may produce different estimates.

For example, suppose we want to estimate the mean height of students at a university. If we draw two different samples of 50 students, we might obtain:

$$\overline{x}_1 = 172.3 \text{ cm} \quad \text{and} \quad \overline{x}_2 = 171.8 \text{ cm}.$$

Which estimate is "correct"? Neither is exactly the true population mean μ, but both are close. How can we quantify this uncertainty?

Confidence Interval Estimation provides a way of quantifying this sampling uncertainty.

A confidence interval for θ is an interval whose endpoints are calculated from the sample and which is constructed so that the procedure has a specified probability of covering the unknown parameter.

A confidence interval can be written in the general form

$$[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)],$$

where L and U are statistics and

$$P_\theta(L \leq \theta \leq U) = 1 - \alpha.$$

The quantity $1 - \alpha$ is called the **confidence level**, while α is called the **risk level** or **significance level of the confidence procedure**.

Thus, point estimation and interval estimation have different objectives:

- **Point estimation** provides a single value;
- **Confidence interval estimation** provides a range of plausible values together with a specified confidence level.

3.2 Confidence Level and Risk Level

Let

$$0 < \alpha < 1.$$

A confidence interval with confidence level $1 - \alpha$ is constructed so that

$$P_\theta\left(L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n)\right) = 1 - \alpha.$$

The quantity α is called the **risk level**.
For example, the commonly used confidence levels are

$$90\%, \qquad 95\%, \qquad 99\%.$$

They correspond respectively to

$$\alpha = 0.10, \qquad \alpha = 0.05, \qquad \alpha = 0.01.$$

Trade-off between Confidence and Precision: For a fixed sample size, increasing the confidence level generally produces a wider interval. Conversely, decreasing the confidence level generally produces a narrower interval.

This illustrates the fundamental compromise: a procedure designed to cover the parameter more frequently generally requires a wider interval.

3.2.1 Interpretation of a Confidence Interval

Suppose that a confidence interval $[L, U]$ has been obtained using a procedure with confidence level $1 - \alpha$.

Frequentist Interpretation: The confidence level is a property of the **procedure** used to construct the interval. In the frequentist framework, the parameter θ is fixed, whereas the interval endpoints are random before the sample is observed.

Thus, the statement

$$P_\theta(L \leq \theta \leq U) = 1 - \alpha$$

refers to the **long-run behavior** of the confidence interval procedure over repeated samples.

Example: If we construct 95% confidence intervals from 100 different samples, we expect approximately 95 of them to contain the true parameter θ.

Important: After the sample has been observed, the resulting interval is fixed. We cannot say "there is a 95% probability that θ is in $[L, U]$" because θ is fixed and the interval is now fixed. The correct statement is: "We are 95% confident that the interval $[L, U]$ contains θ," which refers to the reliability of the procedure.

3.3 Construction of Confidence Intervals

The construction of a confidence interval generally relies on a statistic whose sampling distribution is known or can be determined.

The basic principle is to find a quantity involving the unknown parameter whose distribution does not depend on that parameter, or whose distribution is known after standardization.

Definition 3.1 (Pivotal Quantity)**.** A **pivotal quantity** (or **pivot**) is a function $Q(X_1, \ldots, X_n, \theta)$ of the sample and the unknown parameter θ whose probability distribution does *not* depend on θ or any other unknown parameter.

Example: If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known, then

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

is a pivotal quantity for μ because its distribution is standard normal, which does not depend on μ.

The standard method for constructing a confidence interval relies on finding a pivotal quantity. If we can find constants a and b such that

$$P_\theta(a \leq Q(X, \theta) \leq b) = 1 - \alpha,$$

we can algebraically rearrange the inequalities to isolate θ, yielding the confidence interval $[L(X), U(X)]$.

General Procedure:

1. Identify a pivotal quantity $Q(X, \theta)$ whose distribution is known.
2. Find critical values a and b such that $P(a \leq Q \leq b) = 1 - \alpha$.
3. Solve the inequality $a \leq Q(X, \theta) \leq b$ for θ.
4. The resulting interval is the $1 - \alpha$ confidence interval.

3.4 Confidence Interval for a Mean

Let

$$X_1, \ldots, X_n$$

be a random sample from a population with mean μ and variance σ^2.

The parameter of interest is the population mean μ.

The construction of a confidence interval for μ depends on whether the population variance σ^2 is known or unknown.

3.4.1 Known Variance

Assume first that σ^2 is known.

Under the normal model,

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

the sample mean satisfies

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

Therefore,

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

The quantity Z is a pivotal quantity for μ.

Let $z_{\alpha/2}$ denote the positive number satisfying

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2},$$

where $Z \sim \mathcal{N}(0, 1)$.

Then

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Rearranging gives

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Hence the confidence interval for μ is

$$\boxed{\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]}.$$

The margin of error is

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The interval can therefore be written as

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

3.4.2 Unknown Variance

Suppose now that σ^2 is unknown.

For a normal population,

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

define

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

The statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a Student distribution with $n - 1$ degrees of freedom:

$$T \sim t_{n-1}.$$

The quantity T is a pivotal quantity for μ.

Let $t_{n-1,\alpha/2}$ denote the positive critical value satisfying

$$P(T > t_{n-1,\alpha/2}) = \frac{\alpha}{2}.$$

Then

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1 - \alpha.$$

Therefore,

$$P\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Hence the confidence interval is

$$\boxed{\left[\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right]}.$$

The corresponding margin of error is

$$E = t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

3.4.3 Summary for the Mean

For a normal population, the confidence interval for μ is obtained as follows:

Variance	**Confidence interval for μ**
σ^2 known	$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
σ^2 unknown	$\overline{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$

The normal distribution is used when σ is known, whereas the Student distribution is used when σ is unknown under the normal population assumption.

3.5 Confidence Interval for a Variance

Consider a random sample

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

where both μ and σ^2 may be unknown.

The parameter of interest is the population variance σ^2.

The sample variance is

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

For a normal population, the statistic

$$\frac{(n-1)S^2}{\sigma^2}$$

has a chi-square distribution with $n-1$ degrees of freedom:

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

This quantity is a pivotal quantity for σ^2.

Let $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ denote the appropriate chi-square quantiles, where:

- $\chi^2_{n-1,\alpha/2}$ is the **upper** quantile: $P(\chi^2 > \chi^2_{n-1,\alpha/2}) = \alpha/2$
- $\chi^2_{n-1,1-\alpha/2}$ is the **lower** quantile: $P(\chi^2 > \chi^2_{n-1,1-\alpha/2}) = 1 - \alpha/2$

Then

$$P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) = 1 - \alpha.$$

Because σ^2 appears in the denominator, reversing the inequalities when solving for σ^2 gives

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1 - \alpha.$$

Thus, the confidence interval for the population variance is

$$\boxed{\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]}.$$

Remark 3.1. The chi-square distribution is **not symmetric**, which is why the confidence interval for σ^2 is not centered on S^2. The interval is skewed to the right, reflecting the asymmetry of the chi-square distribution.

This construction relies on the normality assumption. In particular, the chi-square distribution of the standardized sample variance is a specific property of the normal model.

3.6 Confidence Interval for a Proportion

Suppose that a population is divided into two categories and let p denote the proportion of individuals belonging to the category of interest.

Let

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p).$$

The sample proportion is

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

Since

$$\mathbb{E}[\widehat{p}] = p \quad \text{and} \quad \text{Var}(\widehat{p}) = \frac{p(1-p)}{n},$$

the standardized statistic can be approximated by a standard normal variable for sufficiently large samples:

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

Validity Condition: The normal approximation is considered valid when

$$n\widehat{p} \geq 5 \quad \text{and} \quad n(1-\widehat{p}) \geq 5.$$

Using the normal approximation, a confidence interval for p can be constructed. Replacing the unknown p in the standard error by $\widehat{p}$ gives the usual approximate interval (Wald interval)

$$\boxed{\left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]}.$$

Equivalently,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Remark 3.2. The interval constructed above is known as the **Wald interval**. For small sample sizes or proportions near 0 or 1, this interval can produce lower bounds less than 0 or upper bounds greater than 1, and its actual coverage probability may fall below the nominal $1-\alpha$ level. In such cases, alternative methods (such as the Wilson score interval or the Agresti–Coull interval) are recommended in modern statistical practice.

3.7 Confidence Intervals for Two Populations

In many applications, one is interested in comparing two populations. The construction of confidence intervals for the difference of two parameters follows the same pivotal-quantity principle as in the one-sample case.

3.7.1 Difference of Two Means (Known Variances)

Consider two independent random samples

$$X_1, \dots, X_{n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2), \qquad Y_1, \dots, Y_{n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2).$$

Assume that σ_1^2 and σ_2^2 are known. The parameter of interest is $\mu_1 - \mu_2$.

Since

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

the pivotal quantity is

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0, 1).$$

Hence the $1 - \alpha$ confidence interval for $\mu_1 - \mu_2$ is

$$\boxed{\left[(\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]}.$$

3.7.2 Difference of Two Means (Unknown Equal Variances)

Suppose now that σ_1^2 and σ_2^2 are unknown but assumed to be equal: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

The pooled variance estimator is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The pivotal quantity is

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

Hence the $1 - \alpha$ confidence interval for $\mu_1 - \mu_2$ is

$$\boxed{\left[(\overline{X} - \overline{Y}) - t_{n_1+n_2-2,\alpha/2}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\overline{X} - \overline{Y}) + t_{n_1+n_2-2,\alpha/2}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]}.$$

Remark 3.3 (Pooled t-interval). This procedure is also called the **pooled t-interval** because it uses the pooled variance estimator S_p^2.

Remark 3.4 (Welch's interval). When the assumption $\sigma_1^2 = \sigma_2^2$ is questionable, one may use **Welch's interval**, which replaces S_p by $\sqrt{S_1^2/n_1 + S_2^2/n_2}$ and uses a Student distribution with degrees of freedom given by the Welch–Satterthwaite equation. This is the default procedure in most modern statistical software (R, Python).

3.7.3 Ratio of Two Variances

Consider two independent normal samples as above. The parameter of interest is σ_1^2/σ_2^2.

Since

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1},$$

the quantity F is a pivotal quantity for σ_1^2/σ_2^2.

Let $F_{n_1-1,n_2-1,\alpha/2}$ and $F_{n_1-1,n_2-1,1-\alpha/2}$ denote the appropriate F quantiles, where:

- $F_{n_1-1,n_2-1,\alpha/2}$ is the **upper** quantile: $P(F > F_{n_1-1,n_2-1,\alpha/2}) = \alpha/2$
- $F_{n_1-1,n_2-1,1-\alpha/2}$ is the **lower** quantile: $P(F > F_{n_1-1,n_2-1,1-\alpha/2}) = 1 - \alpha/2$

Hence the $1-\alpha$ confidence interval for σ_1^2/σ_2^2 is

$$\boxed{\left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1,n_2-1,\alpha/2}},\ \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right]}.$$

3.7.4 Difference of Two Proportions

Let $\widehat{p}_1$ and $\widehat{p}_2$ be the sample proportions from two independent populations, with true proportions p_1 and p_2.

For sufficiently large sample sizes, the pivotal quantity

$$Z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\widehat{p}_1(1-\widehat{p}_1)/n_1 + \widehat{p}_2(1-\widehat{p}_2)/n_2}} \approx \mathcal{N}(0,1).$$

Hence the approximate $1-\alpha$ confidence interval for $p_1 - p_2$ is

$$\boxed{\left[(\widehat{p}_1 - \widehat{p}_2) - z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}},\ (\widehat{p}_1 - \widehat{p}_2) + z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}\right]}.$$

3.8 Influence of Sample Size on Precision

The sample size has a direct influence on the precision of an estimator and of the corresponding confidence interval.

For the confidence interval for a mean with known variance, the margin of error is

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Thus,

$$E \propto \frac{1}{\sqrt{n}}.$$

For a fixed confidence level and fixed population variance, increasing the sample size decreases the margin of error.

Example: If the sample size is multiplied by four,

$$n \longrightarrow 4n,$$

then

$$\frac{1}{\sqrt{4n}} = \frac{1}{2\sqrt{n}}.$$

Therefore, the margin of error is divided by two.

This illustrates an important property of sampling precision: increasing the sample size improves precision, but the improvement occurs at a rate proportional to the inverse square root of the sample size.

For a proportion, the approximate margin of error is

$$E = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Again, the margin of error decreases as the sample size increases.

The same general principle applies to many confidence interval procedures, although the exact form of the margin of error depends on the parameter and the sampling distribution.

3.9 Determination of the Optimal Sample Size

In many statistical studies, the sample size must be determined before the data are collected.

A common objective is to choose n so that the margin of error does not exceed a prescribed value E at a specified confidence level.

3.9.1 Sample Size for Estimating a Mean

Suppose that the population variance σ^2 is known and that the desired margin of error is E.

For a confidence level $1-\alpha$,

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Solving for n gives

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

Since the sample size must be an integer, the required sample size is

$$\boxed{n = \left\lceil \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 \right\rceil}.$$

The ceiling function ensures that the prescribed margin of error is not exceeded.

If σ is unknown, a preliminary estimate of the standard deviation may be used when planning the study.

3.9.2 Sample Size for Estimating a Proportion

For the approximate confidence interval for a proportion, the margin of error is

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Solving for n gives

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}.$$

Since p is generally unknown before the study, a preliminary estimate $\widetilde{p}$ may be used:

$$n = \frac{z_{\alpha/2}^2 \widetilde{p}(1-\widetilde{p})}{E^2}.$$

When no prior information about p is available, the value $p = 1/2$ is commonly used because $p(1-p) \leq 1/4$ for all $p \in [0,1]$.

This gives the conservative sample size

$$\boxed{n = \left\lceil \frac{z_{\alpha/2}^2}{4E^2} \right\rceil}.$$

This choice guarantees the required sample size for the normal-approximation formula over the possible values of p.

Numerical Example: For a 95% confidence level ($z_{0.025} = 1.96$) and a margin of error $E = 0.03$, the conservative sample size is

$$n = \left\lceil \frac{1.96^2}{4 \times 0.03^2} \right\rceil = \left\lceil \frac{3.8416}{0.0036} \right\rceil = \lceil 1067.1 \rceil = 1068.$$

3.9.3 Factors Affecting the Required Sample Size

The required sample size depends mainly on:

- the confidence level;
- the desired margin of error;
- the variability of the population;
- the parameter being estimated.

For a fixed confidence level, requiring greater precision means choosing a smaller margin of error. Since the sample size is proportional to $1/E^2$, achieving twice the precision in terms of halving the margin of error generally requires approximately four times the sample size.

Similarly, increasing the confidence level increases the critical value and therefore increases the required sample size for a fixed margin of error.

3.10 Conclusion

Confidence interval estimation extends point estimation by incorporating sampling uncertainty into the estimation procedure.

A confidence interval is constructed so that the associated procedure contains the unknown parameter with a specified confidence level.

The chapter considered confidence intervals for fundamental parameters:

- the population mean (one and two samples);
- the population variance and the ratio of two variances;
- the population proportion and the difference of two proportions.

For the population mean, the normal distribution is used when the population variance is known, while the Student distribution is used when the variance is unknown under the normal population model.

For the population variance, the chi-square distribution provides the corresponding confidence interval under normality.

For a population proportion, the normal approximation to the binomial model leads to an approximate confidence interval under appropriate sample-size conditions.

The construction of each interval relies on the identification of a pivotal quantity whose distribution is free of unknown parameters.

Finally, the influence of sample size on precision was examined, together with the determination of a sample size required to achieve a prescribed margin of error at a specified confidence level.

These concepts provide the basis for the construction and interpretation of confidence intervals in practical statistical inference.

Chapter 4

Introduction to Statistical Tests

4.1 Objectives of Statistical Tests

Statistical inference is not limited to estimating unknown population parameters. In many applications, one also needs to make decisions about population characteristics or about assumptions concerning a statistical model.

A **statistical test** provides a formal procedure for using sample information to evaluate a statistical hypothesis.

The general objective of a statistical test is to determine whether the observed data provide sufficient evidence to reject a specified hypothesis about the population.

Important: A statistical test does not prove a hypothesis to be true or false in an absolute sense. It provides a decision based on the information contained in the sample and on a prescribed decision rule.

The main elements of a statistical test are:

- the formulation of hypotheses;
- the choice of a test statistic;
- the specification of a significance level;
- the determination of a critical region;
- the calculation of the observed value of the test statistic;
- the statistical decision;
- the interpretation of the decision in the context of the problem.

4.2 Null and Alternative Hypotheses

A statistical test begins with the formulation of two hypotheses.

The **null hypothesis**, denoted by H_0, represents the hypothesis that is tested directly.

The **alternative hypothesis**, denoted by H_1 or sometimes H_a, represents the competing hypothesis.

Example: Suppose that μ denotes a population mean and that a reference value μ_0 is given. One may consider

$$H_0 : \mu = \mu_0$$

against

$$H_1 : \mu \neq \mu_0.$$

The alternative hypothesis determines the form of the statistical test.

4.2.1 Two-Sided Alternative

A two-sided test is used when deviations from the null value in either direction are of interest.

For example,

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0.$$

Evidence against H_0 may then arise from values of the test statistic that are either sufficiently small or sufficiently large.

4.2.2 Right-Sided Alternative

A right-sided test is used when values larger than the null value are of interest.

The hypotheses take the form

$$H_0 : \theta \leq \theta_0 \quad \text{against} \quad H_1 : \theta > \theta_0.$$

In some parametric settings, the null hypothesis may be written simply as $H_0 : \theta = \theta_0$ when the test is constructed against the specified one-sided alternative.

4.2.3 Left-Sided Alternative

A left-sided test is used when values smaller than the null value are of interest:

$$H_0 : \theta \geq \theta_0 \quad \text{against} \quad H_1 : \theta < \theta_0.$$

The direction of the alternative hypothesis determines the location of the critical region.

4.3 Types of Statistical Tests

Statistical tests can be classified according to the hypotheses and the parameter being studied.

For a scalar parameter θ, the three basic forms are:

- **Two-sided test:** $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$;
- **Right-sided test:** $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$;
- **Left-sided test:** $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$.

Statistical tests may also be classified according to the statistical model and the parameter under investigation.

In this course, the principal parametric tests concern:

- a population mean;
- a population variance;
- a population proportion;
- the difference between two population means;
- the comparison of two population variances;
- the difference between two population proportions.

The appropriate test depends on the assumptions concerning the population and on whether the population variance is known or unknown.

4.4 Neyman–Pearson Principle

The Neyman–Pearson principle provides a fundamental framework for the construction of statistical tests.

Consider testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$.

Suppose that the observations have density or probability mass function $f(x; \theta)$.

The likelihood ratio compares the likelihood under the null hypothesis with the likelihood under the alternative hypothesis:

$$\Lambda(x) = \frac{f(x; \theta_0)}{f(x; \theta_1)}.$$

A likelihood-ratio test rejects H_0 when $\Lambda(x)$ is sufficiently small, or equivalently when

$$\frac{f(x; \theta_1)}{f(x; \theta_0)}$$

is sufficiently large.

The threshold is chosen so that the probability of rejecting H_0 when H_0 is true is controlled at the prescribed significance level.

Neyman–Pearson Lemma: The likelihood-ratio test provides the most powerful test of a given size among the corresponding class of tests.

This principle is fundamental for understanding the construction of statistical tests.

4.5 Critical and Acceptance Regions

Let $T = T(X_1, \ldots, X_n)$ be a test statistic.

A statistical test divides the possible values of the test statistic into two regions.

The **critical region** (also called the rejection region), denoted by $\mathcal{C}$, is the set of values for which the null hypothesis is rejected.

The **acceptance region** is the complementary region in which the null hypothesis is not rejected.

The decision rule is:

$$\begin{cases} \text{Reject } H_0, & T \in \mathcal{C}, \\ \text{Do not reject } H_0, & T \notin \mathcal{C}. \end{cases}$$

The critical region depends on the form of the alternative hypothesis:

- For a **right-sided test**, the critical region generally lies in the upper tail of the distribution of the test statistic.

- For a **left-sided test**, it lies in the lower tail.

- For a **two-sided test**, the critical region is generally divided between the two tails.

4.6 Significance Level

The **significance level** of a test is denoted by α, with $0 < \alpha < 1$.

It is the probability of rejecting the null hypothesis when the null hypothesis is true.

For a simple null hypothesis, this can be written as

$$\alpha = P_{\theta_0}(T \in \mathcal{C}).$$

For a composite null hypothesis, the size of the test is generally defined through the largest rejection probability over the parameter values allowed under H_0.

Common significance levels are

$$0.10, \qquad 0.05, \qquad 0.01.$$

The significance level is fixed **before** applying the test.

A smaller value of α requires stronger evidence against H_0 before the null hypothesis is rejected.

4.7 Type I and Type II Errors

A statistical decision can lead to two types of errors.

A **Type I error** occurs when the null hypothesis is rejected even though it is true.

The probability of a Type I error is controlled by the significance level α.

A **Type II error** occurs when the null hypothesis is not rejected even though the alternative hypothesis is true.

For a particular alternative parameter value θ, the probability of a Type II error is denoted by $\beta(\theta)$.

Thus,

$$\beta(\theta) = P_\theta(T \notin \mathcal{C}),$$

for parameter values belonging to the alternative hypothesis.

The two errors are conceptually different:

- **Type I error:** rejecting a true null hypothesis;

- **Type II error:** not rejecting a false null hypothesis.

In general, reducing the probability of a Type I error while keeping the sample size fixed can increase the probability of a Type II error.

4.8 Power Function of a Test

The **power function** of a statistical test, denoted by $\pi(\theta)$, is defined for *all* $\theta \in \Theta$ as the probability of rejecting the null hypothesis:

$$\pi(\theta) = P_\theta(T \in \mathcal{C}).$$

For parameter values in the alternative hypothesis ($\theta \in H_1$), the power is $\pi(\theta) = 1 - \beta(\theta)$, representing the probability of correctly rejecting a false null hypothesis.

For parameter values in the null hypothesis ($\theta \in H_0$), the power function gives the probability of a Type I error, which is bounded by the significance level α.

Properties of a Good Test: A good test has a power function that is:

- close to α (or less) for $\theta \in H_0$;
- rapidly approaching 1 for $\theta \in H_1$.

The power depends on the parameter value, the significance level, the sample size, the test statistic, and the underlying statistical model.

4.9 The p-value Approach

While the critical region approach requires pre-specifying α, modern statistical practice heavily relies on the **p-value**.

Definition 4.1. The **p-value** is the probability, computed under the assumption that the null hypothesis H_0 is true, of obtaining a test statistic value at least as extreme as the one actually observed.

The p-value provides a continuous measure of the evidence against H_0:

- A **small p-value** (typically $\leq \alpha$) indicates strong evidence against H_0, leading to its rejection.
- A **large p-value** ($> \alpha$) indicates weak evidence against H_0, leading to the decision "do not reject H_0".

The decision rule can be succinctly stated as: **Reject** H_0 **if p-value** $\leq \alpha$.

4.9.1 Calculation of the p-value

The exact form of the p-value depends on the alternative hypothesis:

- For a **right-sided test** ($H_1 : \theta > \theta_0$):

$$\text{p-value} = P_{\theta_0}(T \geq t_{\text{obs}}).$$

- For a **left-sided test** ($H_1 : \theta < \theta_0$):

$$\text{p-value} = P_{\theta_0}(T \leq t_{\text{obs}}).$$

- For a **two-sided test** ($H_1 : \theta \neq \theta_0$):

$$\text{p-value} = 2 \times \min\left\{P_{\theta_0}(T \geq t_{\text{obs}}),\, P_{\theta_0}(T \leq t_{\text{obs}})\right\}.$$

The advantage of the p-value approach is that it provides more information than a simple reject/do-not-reject decision at a fixed α level.

4.10 Decision Table

The possible decisions of a statistical test can be summarized as follows:

	H_0 true	H_1 true
Reject H_0	Type I error (probability α)	Correct decision (probability $1-\beta$)
Do not reject H_0	Correct decision (probability $1-\alpha$)	Type II error (probability β)

The probability of a Type I error is controlled by α, whereas the probability of a Type II error is represented by $\beta(\theta)$ for a specific alternative parameter value.

The **power** is $1-\beta(\theta)$.

This table emphasizes that a statistical decision is made under uncertainty. The test provides a rule for making the decision based on the sample.

4.11 General Steps of a Statistical Test

A statistical test can be carried out through the following general steps:

Step 1: Formulate the hypotheses
Specify the null and alternative hypotheses: H_0 and H_1.

Step 2: Choose the significance level
Select a value α before examining the result of the test.

Step 3: Choose the test statistic
Construct a statistic whose sampling distribution is known, or can be determined or approximated, under the null hypothesis.

Step 4: Determine the critical region
Use the distribution of the test statistic under H_0 to determine the critical value or critical values corresponding to the chosen significance level.

Step 5: Calculate the observed test statistic
Use the sample observations to calculate the numerical value of the test statistic.

Step 6: Make the statistical decision
If the observed value belongs to the critical region, reject H_0. Otherwise, do not reject H_0.

Alternatively, compute the p-value and compare it with α: reject H_0 if p-value $\leq \alpha$.

Step 7: State the conclusion
The statistical conclusion should be expressed in the context of the original problem.

Important: It is important to distinguish between "reject H_0" and "do not reject H_0". Failure to reject H_0 does not establish that H_0 is true.

4.12 Standard Statistical Tables

The construction of many classical statistical tests requires critical values from standard probability distributions.

The main distributions used in this course are:

- the standard normal distribution $\mathcal{N}(0,1)$;
- Student's t distribution;
- the chi-square distribution χ^2;
- the Fisher–Snedecor distribution F.

4.12.1 Normal Distribution

If $Z \sim \mathcal{N}(0,1)$, then critical values can be obtained from the standard normal distribution table.

For a two-sided test at significance level α, the critical values are $-z_{\alpha/2}$ and $z_{\alpha/2}$.

For a right-sided test, the critical value is z_α.

For a left-sided test, the corresponding lower critical value is $-z_\alpha$.

4.12.2 Student's t Distribution

The Student's t distribution depends on the number of degrees of freedom.

It is used, in particular, for tests concerning a population mean when the population variance is unknown and the normal population assumption is appropriate.

The critical values depend on both the significance level and the degrees of freedom.

4.12.3 Chi-Square Distribution

The chi-square distribution depends on its degrees of freedom.

It is used in particular for tests concerning a population variance under the normal population assumption.

If $X \sim \chi^2_\nu$, then its distribution is determined by the degrees of freedom ν.

4.12.4 Fisher–Snedecor Distribution

The Fisher–Snedecor distribution, usually denoted by F, depends on two degrees of freedom.

It is used in particular for comparing two population variances under the appropriate normality and independence assumptions.

4.13 Parametric Tests

Parametric tests are based on a specified statistical model involving one or more unknown parameters.

The tests considered in this section concern means, variances, proportions, and comparisons between two populations.

4.14 Test for a Population Mean

Let $X_1, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2.

Suppose that the objective is to test $H_0 : \mu = \mu_0$.

4.14.1 Known Variance

Suppose that σ^2 is known.

Test statistic:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Under H_0, $Z \sim \mathcal{N}(0,1)$.

Critical regions:

- For the two-sided alternative $H_1 : \mu \neq \mu_0$: $|Z| > z_{\alpha/2}$
- For the right-sided alternative $H_1 : \mu > \mu_0$: $Z > z_\alpha$
- For the left-sided alternative $H_1 : \mu < \mu_0$: $Z < -z_\alpha$

Conditions of validity: The population is normally distributed, or the sample size is sufficiently large ($n \geq 30$) for the Central Limit Theorem to apply.

4.14.2 Unknown Variance

Suppose that σ^2 is unknown.

Test statistic:

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

Under H_0, $T \sim t_{n-1}$.

Critical regions:

- For the two-sided alternative $H_1 : \mu \neq \mu_0$: $|T| > t_{n-1,\alpha/2}$
- For the right-sided alternative $H_1 : \mu > \mu_0$: $T > t_{n-1,\alpha}$
- For the left-sided alternative $H_1 : \mu < \mu_0$: $T < -t_{n-1,\alpha}$

Conditions of validity: The population is normally distributed.

4.15 Test for a Population Variance

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

Suppose that the objective is to test $H_0 : \sigma^2 = \sigma_0^2$.

Test statistic:

$$Q = \frac{(n-1)S^2}{\sigma_0^2}.$$

Under H_0, $Q \sim \chi^2_{n-1}$.

Critical regions (CORRIGÉES):

- For the two-sided alternative $H_1 : \sigma^2 \neq \sigma_0^2$:

$$Q < \chi^2_{n-1,1-\alpha/2} \quad \text{or} \quad Q > \chi^2_{n-1,\alpha/2}$$

- For the right-sided alternative $H_1 : \sigma^2 > \sigma_0^2$:
$$Q > \chi^2_{n-1,\alpha}$$
- For the left-sided alternative $H_1 : \sigma^2 < \sigma_0^2$:
$$Q < \chi^2_{n-1,1-\alpha}$$

Conditions of validity: The population is normally distributed. The use of the chi-square distribution here relies on the normality assumption.

4.16 Test for a Population Proportion

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, and let $\widehat{p}$ denote the sample proportion:
$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Suppose that we want to test $H_0 : p = p_0$.

Test statistic:
$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Under H_0, $Z \approx \mathcal{N}(0,1)$ for sufficiently large samples.

Critical regions:

- For the two-sided alternative $H_1 : p \neq p_0$: $|Z| > z_{\alpha/2}$
- For $H_1 : p > p_0$: $Z > z_\alpha$
- For $H_1 : p < p_0$: $Z < -z_\alpha$

Conditions of validity: The sample size is sufficiently large, specifically $np_0 \geq 5$ and $n(1-p_0) \geq 5$.

4.17 Comparison of Two Population Means

Suppose that two independent populations have means μ_1 and μ_2.

Let $X_1, \ldots, X_{n_1}$ be a sample from the first population and $Y_1, \ldots, Y_{n_2}$ a sample from the second population.

The parameter of interest is $\mu_1 - \mu_2$.

A common hypothesis is $H_0 : \mu_1 - \mu_2 = 0$, which is equivalent to $H_0 : \mu_1 = \mu_2$.

4.17.1 Known Variances

Suppose that the two population variances σ_1^2 and σ_2^2 are known.

Test statistic:
$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Under the null hypothesis, $Z \sim \mathcal{N}(0,1)$.

For $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, the two-sided critical region is $|Z| > z_{\alpha/2}$.

4.17.2 Unknown Variances

When the variances are unknown, the test procedure depends on the assumptions made about the two populations.

Case 1: Equal variances ($\sigma_1^2 = \sigma_2^2$)
The pooled variance estimator is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The test statistic is

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Under the null hypothesis, $T \sim t_{n_1+n_2-2}$.

The critical region is then determined from the Student distribution according to the chosen alternative hypothesis.

Remark 4.1 (Welch's t-test). The pooled-variance t-test assumes $\sigma_1^2 = \sigma_2^2$. When this assumption is questionable or known to be false, it is standard practice to use **Welch's t-test**, which does not assume equal variances. The test statistic is

$$T_W = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

and its distribution is approximated by a Student's t-distribution with degrees of freedom given by the Welch–Satterthwaite equation:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

This is the default two-sample t-test in most modern statistical software (e.g., R, Python).

4.18 Comparison of Two Population Variances (CORRIGÉ ET ENRICHI)

Suppose that two independent normal populations have variances σ_1^2 and σ_2^2.

The hypothesis of equal variances is $H_0 : \sigma_1^2 = \sigma_2^2$.

Test statistic:

$$F = \frac{S_1^2}{S_2^2}.$$

Under the null hypothesis, $F \sim F_{n_1-1,n_2-1}$.

Critical regions (FORMULES EXPLICITES):

- For a two-sided alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$:

$$F < F_{n_1-1,n_2-1,1-\alpha/2} \quad \text{or} \quad F > F_{n_1-1,n_2-1,\alpha/2}$$

- For a right-sided alternative $H_1 : \sigma_1^2 > \sigma_2^2$:

$$F > F_{n_1-1,n_2-1,\alpha}$$

- For a left-sided alternative $H_1 : \sigma_1^2 < \sigma_2^2$:

$$F < F_{n_1-1,n_2-1,1-\alpha}$$

Conditions of validity: The procedure relies on the independence and normality assumptions for the two samples.

4.19 Comparison of Two Population Proportions

Let

$$\widehat{p}_1 = \frac{X_1}{n_1}, \qquad \widehat{p}_2 = \frac{X_2}{n_2}$$

be the sample proportions from two independent populations.

The parameter of interest is $p_1 - p_2$.

To test $H_0 : p_1 = p_2$, a pooled estimate of the common proportion is used:

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}.$$

Test statistic:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Under the null hypothesis and suitable large-sample conditions, $Z \approx \mathcal{N}(0,1)$.

Critical regions:

- For the two-sided alternative $H_1 : p_1 \neq p_2$: $|Z| > z_{\alpha/2}$
- For $H_1 : p_1 > p_2$: $Z > z_\alpha$
- For $H_1 : p_1 < p_2$: $Z < -z_\alpha$

Conditions of validity: The normal approximation should only be used when the sample sizes and observed counts are sufficiently large for the approximation to be appropriate (typically $n_1\widehat{p}_1 \geq 5$, $n_1(1-\widehat{p}_1) \geq 5$, $n_2\widehat{p}_2 \geq 5$, and $n_2(1-\widehat{p}_2) \geq 5$).

4.20 Illustrative Example: Test for a Population Mean

Context: A manufacturer claims that the mean lifetime of a certain type of light bulb is 1000 hours. A consumer association wants to test this claim.

Data: A random sample of $n = 25$ light bulbs is tested, yielding a sample mean $\overline{x} = 980$ hours and a sample standard deviation $s = 50$ hours.

Step 1: Formulate the hypotheses

$$H_0 : \mu = 1000 \quad \text{versus} \quad H_1 : \mu \neq 1000.$$

Step 2: Choose the significance level
Let $\alpha = 0.05$.

Step 3: Choose the test statistic
Since the population variance is unknown and the sample size is small ($n < 30$), we use the Student t-test:

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

Under H_0, $T \sim t_{n-1} = t_{24}$.

Step 4: Determine the critical region
For a two-sided test at $\alpha = 0.05$ with 24 degrees of freedom:

$$t_{24,0.025} = 2.064.$$

The critical region is $|T| > 2.064$.

Step 5: Calculate the observed test statistic

$$t_{\text{obs}} = \frac{980 - 1000}{50/\sqrt{25}} = \frac{-20}{10} = -2.0.$$

Step 6: Make the statistical decision
Since $|t_{\text{obs}}| = 2.0 < 2.064$, we do **not reject** H_0.

Step 7: State the conclusion
At the 5% significance level, there is not enough evidence to conclude that the mean lifetime of the light bulbs differs from 1000 hours.

Alternative approach using p-value:
The p-value is $2 \times P(T_{24} \leq -2.0) \approx 2 \times 0.028 = 0.056$.
Since p-value $= 0.056 > 0.05 = \alpha$, we do not reject H_0 (same conclusion).

4.21 Duality Between Tests and Confidence Intervals

There is a fundamental connection between hypothesis testing and confidence interval estimation.

Theorem: A two-sided test of $H_0 : \theta = \theta_0$ at significance level α rejects H_0 if and only if θ_0 does not belong to the $1 - \alpha$ confidence interval for θ.

Example: Suppose we want to test $H_0 : \mu = 100$ at $\alpha = 0.05$. If the 95% confidence interval for μ is $[95, 105]$, then since $100 \in [95, 105]$, we do not reject H_0.

This duality provides a useful way to interpret the results of statistical tests in terms of plausible values for the parameter.

4.22 Summary of Parametric Tests (ENRICHI)

The principal parametric tests introduced in this chapter can be summarized as follows:

Parameter	Test statistic	Distribution	Conditions
μ, σ^2 known	$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$\mathcal{N}(0,1)$	Normal or $n \geq 30$
μ, σ^2 unknown	$\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$	t_{n-1}	Normal population
σ^2	$\frac{(n-1)S^2}{\sigma_0^2}$	χ^2_{n-1}	Normal population
p	$\frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$	$\mathcal{N}(0,1)$ approx.	$np_0 \geq 5$, $n(1-p_0) \geq 5$

For two populations, the corresponding procedures involve the normal, Student, Fisher–Snedecor, or normal approximation distributions depending on the parameter and assumptions.

4.23 Conclusion

Statistical testing provides a formal framework for making decisions about population characteristics using sample information.

The construction of a statistical test begins with the formulation of a null hypothesis and an alternative hypothesis. The significance level determines the allowed probability of a Type I error, while the power function of the test describes its ability to detect departures from the null hypothesis across the entire parameter space.

The Neyman–Pearson principle provides a fundamental basis for constructing powerful tests in the simple-versus-simple setting.

The concepts of critical region, acceptance region, Type I error, Type II error, power, and p-value are central to the interpretation of statistical tests.

The chapter also introduced the main probability distributions used in classical parametric testing: the normal, Student, chi-square, and Fisher–Snedecor distributions.

Finally, tests concerning a population mean, variance, and proportion were presented, together with procedures for comparing two population means, variances, and proportions, including the important Welch's t-test for the comparison of two means with unequal variances.

These methods provide the basic tools required for applying statistical hypothesis testing to common inference problems.