الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
Tasdawit Akli Muḥend Ulḥağ - Tubirett -
Faculté des Sciences de la Nature
et de la Vie et des Sciences de la Terre

وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
- البويرة -
كلية علوم الطبيعة والحياة وعلوم الأرض

M. TABCHOUCHE

Maitre de conférences en mathématiques

Département des Sciences Agronomiques, FSNVST-UAMOB

Mobile : +213 675674858

Mail : n.tabchouche@univ-bouira.dz

Ouvrage pédagogique

Biostatistiques

Destiné aux étudiants

Niveau : Deuxième Année Licence.
Spécialité : Ingénieurs en Agronomie.
Semestre : 01
Nature de l'unité d'enseignement : UEM
Crédit : 02
Coefficient : 02
Volume horaire semestriel : 45h

Année universitaire : 2025-2026

Foreword

This course in Biostatistics for Agronomy Engineers, prepared for the academic year 2025/2026 and realized by Dr. Nesrine Tabchouche, was designed to help students think with data and act with scientific care. Agronomic questions arise in conditions that are complex and often uncertain. Decisions about varieties, soils, pests, and resources require methods that quantify uncertainty, test clear hypotheses, and communicate results in language that is both precise and useful for practice.

The material proceeds from foundations to application. It begins with the core vocabulary of statistics, then develops estimation and confidence intervals, and finally introduces the full suite of hypothesis tests that students will need in field and laboratory contexts. Topics include comparisons of means through t procedures and ANOVA, analysis of proportions and categorical data, and linear regression for relationships between variables. Each chapter links formulae to interpretation, with worked examples that reflect real agronomic settings.

The pedagogical aim is professional competence. Students are encouraged to plan analyses before seeing outcomes, when possible, to check assumptions with diagnostic tools, to report effect sizes together with exact p-values and confidence intervals, and to write results that decision makers can understand. Reproducibility is treated as a habit, not an afterthought. Data preparation and code should be documented with enough clarity that another researcher can verify and extend the work.

These notes are intended to be used actively. Read the statements of assumptions, attempt the exercises before consulting solutions, and compare your reasoning with the reporting templates that follow each major test. When uncertainty persists, treat it as part of responsible scientific practice and state its implications for the conclusion.

I am grateful to the students whose questions sharpened the presentation, to colleagues who offered feedback on examples and notation, and to the staff who supported the logistics of teaching and assessment. Any remaining errors are my own. I invite suggestions that can improve future editions.

Academic Year 2025/2026

Dr. RAI-TBCHOUCHE Nesrine

List of Figures

List of Tables

List of Abbreviations

- SD – Standard Deviation
- SE – Standard Error
- CI – Confidence Interval
- CV – Coefficient of Variation
- IQR – Interquartile Range
- df – Degrees of Freedom
- ANOVA – Analysis of Variance

Hypothesis testing

- H_0 – Null Hypothesis
- H_1 or H_a – Alternative Hypothesis
- p-value (probability value)
- α – Significance Level (Type I error rate)
- β – Type II Error Rate

Correlation and regression

- r – Pearson Correlation Coefficient
- ρ – Spearman Rank Correlation
- R^2 – Coefficient of Determination

Data and sampling

- N – Population size
- n – Sample size
- CLT – Central Limit Theorem

Contents

Listings

Introduction

Prepared for the academic year 2025/2026 and realized by Dr. Nesrine TABCHOUCHE, this second-year course in Biostatistics for Agronomy Engineers develops the habits of mind required to turn field and laboratory data into trustworthy conclusions. Agronomy increasingly depends on quantitative evidence. Variety trials, soil and water assessments, plant health surveillance, and resource allocation all require tools that can measure uncertainty, test plausible explanations, and communicate findings in a clear and responsible way.

The course begins with the language of data. Students review populations and samples, measurement scales, and the distinction between parameters and statistics. Core distributions such as the normal, binomial, and Poisson are presented with intuition, formulae, and examples that arise naturally in agricultural research. Descriptive summaries are linked to the study of shape and spread through quartiles, percentiles, skewness, and kurtosis, which helps students read datasets with nuance rather than relying only on averages.

Building on these foundations, the course turns to inference. Point estimation and confidence intervals are introduced as complementary ways to quantify what the data suggest about unknown quantities. Hypothesis testing then provides a disciplined approach to decision making, with careful attention to assumptions, power, and effect sizes. Students learn the appropriate use of z and t procedures for means, chi-square methods for categorical data, F tests and ANOVA for comparing several means, and the logic of post-hoc comparisons. The final part of the course consolidates bivariate and multivariate analysis through correlation, simple linear regression, scatter-plot interpretation, and multiple regression with an emphasis on model diagnostics.

Throughout, learning is anchored in agronomic applications. Exercises use real or realistic datasets so that students repeatedly practice good analytical workflow: pre-specifying questions, checking assumptions, reporting exact p-values alongside confidence intervals, and interpreting results in terms of agronomic impact such as yield, risk, and efficiency. By the end of the semester, students should be able to select appropriate methods, implement them correctly, and present results that can support responsible decisions in research and practice.

Chapter 1

Basic Concepts

1.1 Population and Sample [12, (Lawal, 2014)]

- **Population:** The entire set of individuals or items that share at least one characteristic of interest. It is the group we want to study or draw conclusions about.
- **Sample:** A subset of the population selected for observation and analysis. It should be representative of the population.

1.2 Variable

- A **variable** is any characteristic or attribute that can take different values.
- **Types of variables:**
  - **Quantitative:** Numeric values (e.g., height, weight, age).
    - *Discrete:* Countable values (e.g., number of children).
    - *Continuous:* Any value within a range (e.g., temperature).
  - **Qualitative (Categorical):** Non-numeric values (e.g., color, gender).
    - *Nominal:* No natural order (e.g., blood type).
    - *Ordinal:* Natural order exists (e.g., education level).

1.3 Parameter and Statistic

- **Parameter:** A numerical summary describing a population (e.g., population mean μ).
- **Statistic:** A numerical summary describing a sample (e.g., sample mean $\bar{x}$).

1.4 Descriptive and Inferential Statistics

*

- **Descriptive Statistics:** Methods for summarizing and describing the important characteristics of a dataset.
- **Inferential Statistics:** Techniques for drawing conclusions or making decisions about a population based on sample data.

1.5 Distributions [12, (Lawal, 2014)]

1.5.1 Normal Distribution

The normal distribution is a continuous probability distribution that is symmetric and bell-shaped.

- **Notation:** $X \sim \mathcal{N}(\mu, \sigma^2)$
- **Probability density function (PDF):**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Properties:**
  - Mean = μ
  - Variance = σ^2
  - Symmetric around the mean
  - 68%, 95%, 99.7% rule (empirical rule)
- **Standard normal:** $\mathcal{N}(0,1)$ with $Z = \frac{X-\mu}{\sigma}$

1.5.2 Binomial Distribution

The binomial distribution is a discrete distribution modeling the number of successes in n independent Bernoulli trials.

- **Notation:** $X \sim \text{Bin}(n, p)$
- **Probability mass function (PMF):**

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

- **Properties:**
  - Mean = np
  - Variance = $np(1-p)$
  - Skewed if p is far from 0.5
- **Applications:** Coin tosses, success/failure experiments

1.5.3 Poisson Distribution

The Poisson distribution is a discrete distribution modeling the number of events in a fixed interval of time or space, assuming events occur independently and at a constant rate.

- **Notation:** $X \sim \text{Poisson}(\lambda)$
- **Probability mass function (PMF):**

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

- **Properties:**
  - Mean = Variance = λ
  - Skewed for small λ, approaches normal for large λ
- **Applications:** Number of calls per hour, emails per day, accidents, arrivals

Table 1.1: Comparison of Normal, Binomial, and Poisson Distributions

Feature	**Normal**	**Binomial**	**Poisson**
Type	Continuous	Discrete	Discrete
Domain	$x \in \mathbb{R}$	$k = 0, 1, \ldots, n$	$k = 0, 1, 2, \ldots$
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\text{Bin}(n, p)$	$\text{Poisson}(\lambda)$
PDF / PMF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$\frac{\lambda^k e^{-\lambda}}{k!}$
Mean	μ	np	λ
Variance	σ^2	$np(1-p)$	λ
Symmetry	Symmetric	Skewed if $p \neq 0.5$	Skewed if λ is small
Limiting case	–	Approaches normal as n increases	Approaches normal as λ increases
Typical use	Natural measurements	Repeated success/failure trials	Counting rare events

1.6 Quartiles and Interquartile Range [12, (Lawal, 2014)]

Quartiles divide an ordered data set into four equal parts.

- **First Quartile** (Q_1): The value below which 25% of the data fall. Also called the lower quartile.
- **Second Quartile** (Q_2): The median. It divides the data into two equal halves.
- **Third Quartile** (Q_3): The value below which 75% of the data fall. Also called the upper quartile.

The quartiles can be visualized on a boxplot, which shows:

- Minimum
- Q_1
- Q_2 (Median)
- Q_3
- Maximum

Formula for Grouped Data

$$Q_k = L + \left(\frac{\frac{kN}{4} - F}{f} \right) \cdot h.$$

- $Q_k = k - th$ quartile (Q_1, Q_2, Q_3)
- L=Lower boundary of the quartile class

- N=Total frequency
- F=Cumulative frequency before the quartile class
- f=Frequency of the quartile class
- h=Class width

Interquartile Range (IQR)

The interquartile range is given by:

$$\text{IQR} = Q_3 - Q_1$$

Measures the spread of the middle 50

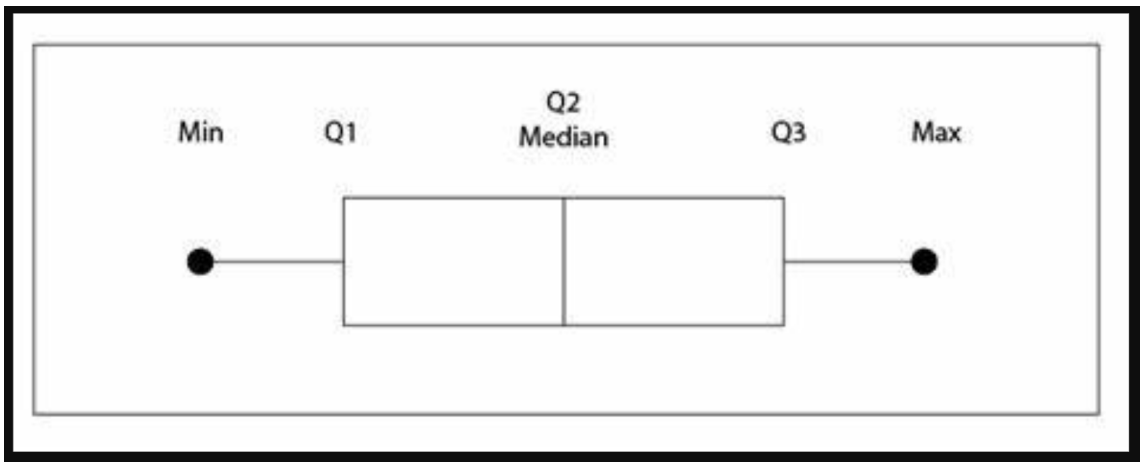


Figure 1.1: The Boxplot

1.7 Deciles and Percentiles [12, (Lawal, 2014)]

1.7.1 Deciles

- **Definition:** Deciles divide a sorted dataset into **10 equal parts**.
- **Notation:** $D_1, D_2, \ldots, D_9$ are the 9 deciles that split the dataset.
- **Interpretation:**
  - D_1: 10% of data is below this value.
  - D_5: 50% of data is below this value (also called the **median**).
  - D_9: 90% of data is below this value.

1.7.2 Percentiles

- **Definition:** Percentiles divide a sorted dataset into **100 equal parts**.
- **Notation:** $P_1, P_2, \ldots, P_{99}$ are the 99 percentiles.
- **Interpretation:**

- P_{25}: 25% of data is below this value (**first quartile** or Q_1).
- P_{50}: 50% of data is below this value (**median** or Q_2).
- P_{75}: 75% of data is below this value (**third quartile** or Q_3).

1.7.3 Formula for Grouped Data

- **Percentile Formula:**

$$P_k = L + \left(\frac{k \cdot N}{100} - F \right) \cdot \frac{h}{f}$$

- **Decile Formula:** (same formula using $k \cdot N/10$)

$$D_k = L + \left(\frac{k \cdot N}{10} - F \right) \cdot \frac{h}{f}$$

- **Where:**
  - L: Lower boundary of the class
  - N: Total number of values
  - F: Cumulative frequency before the class
  - f: Frequency of the class
  - h: Class width
  - k: Desired percentile or decile number

Notes

- Percentiles are useful when detailed classification is needed (e.g., exams, salaries).
- Deciles are used when a broader categorization is sufficient.

1.8 Measures of Central Tendency [20, (Rosner, 2006)]

- **Mean (Average):**
  - **Ungrouped Data** Given a sample of n values $x_1, x_2, \ldots, x_n$:

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

    The arithmetic average of the values in the dataset.
  - **Grouped Data** Suppose data is grouped into classes with midpoints x_i and frequencies f_i, total frequency $N = \sum f_i$:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} f_i x_i$$

- **Median:** The middle value when the data are ordered. If n is even, it is the average of the two middle values.

- **Ungrouped Data**
  - Order the data from smallest to largest.
  - If n is odd, the median is the value at position:

$$Med = x_{\frac{n+1}{2}},$$

  - If n is even, the median is the average of the two middle values:

$$Med = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2},$$

- **Grouped Data**

$$\text{Median} = L + \left(\frac{\frac{N}{2} - F}{f} \right) \times h$$

  - L = lower boundary of the median class
  - N = total frequency
  - F = cumulative frequency before the median class
  - f = frequency of the median class
  - h = class width (interval size)

- **Mode:**
  - **Ungrouped Data**: The value that appears most frequently in the dataset.
  - **Grouped Data** When data is grouped into classes, the mode is estimated using the following formula:

$$\text{Mode} = L + \left(\frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \right) \times h$$

    - L = lower boundary of the modal class
    - f_m = frequency of the modal class (the class with highest frequency)
    - f_1 = frequency of the class before the modal class
    - f_2 = frequency of the class after the modal class
    - h = class width (interval size)

    Steps:

    1. Identify the modal class (class with the highest frequency).
    2. Apply the formula to estimate the mode.

1.9 Measures of Dispersion [20, (Rosner, 2006)]

- **Range:** The difference between the maximum and minimum values:

$$\text{Range} = \max(x_i) - \min(x_i)$$

- **Variance:**
  - **Ungrouped Data**

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 \right) - \bar{x}^2$$

– **Grouped Data**

$$s^2 = \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2$$

A measure of how data points deviate from the mean.

- **Standard Deviation:**

$$s = \sqrt{s^2}$$

The square root of the variance, expresses spread in the same units as the data.

Note

– x_i: Class midpoint

– f_i: Frequency of the i^{th} class

– N: Total number of observations ($N = \sum f_i$)

Histogram

```r
# Define class values
xi <- c(1, 2, 3, 4, 5)

# Define corresponding frequencies
ni <- c(2, 5, 8, 4, 1)

# Expand the data based on frequencies
data <- rep(xi, times = ni)

# Create the histogram
hist(data,
     breaks = seq(0.5, 5.5, by = 1),
     col = "skyblue",
     border = "white",
     main = "Histogram␣from␣(xi,␣ni)",
     xlab = "Class␣Values␣(xi)",
     ylab = "Frequency")
```

Listing 1.1: R code to generate a histogram from (xi, ni)

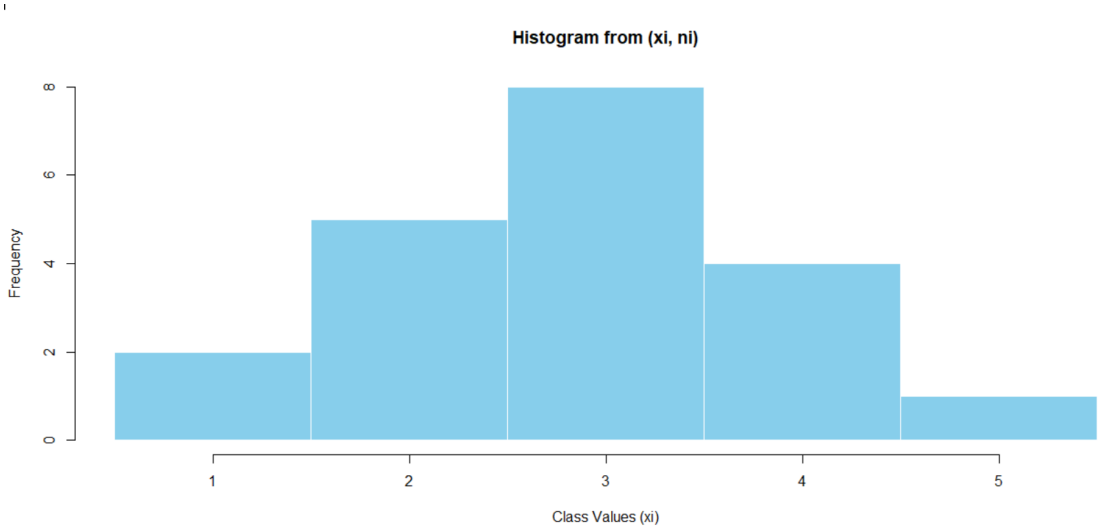


Figure 1.2: The histogram from (xi, ni)

1.10 Skewness and Kurtosis [9, (Hatem et all,. 2022)]

1.10.1 Skewness

Skewness measures the asymmetry of a distribution around its mean.

- If **skewness** $= 0$, the distribution is symmetric.
- If **skewness** > 0, the distribution is positively skewed (right tail longer).
- If **skewness** < 0, the distribution is negatively skewed (left tail longer).

Formula for sample skewness:

$$\text{Skewness} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

where:

- n is the number of observations,
- $\bar{x}$ is the sample mean,
- s is the sample standard deviation.

1.10.2 Kurtosis (Flattening)

Kurtosis indicates whether a distribution is more or less peaked than a normal distribution.

- If **kurtosis** $= 3$, the distribution is **mesokurtic** (normal-like).
- If **kurtosis** > 3, it is **leptokurtic** (sharper peak, heavy tails).
- If **kurtosis** < 3, it is **platykurtic** (flatter peak, light tails).

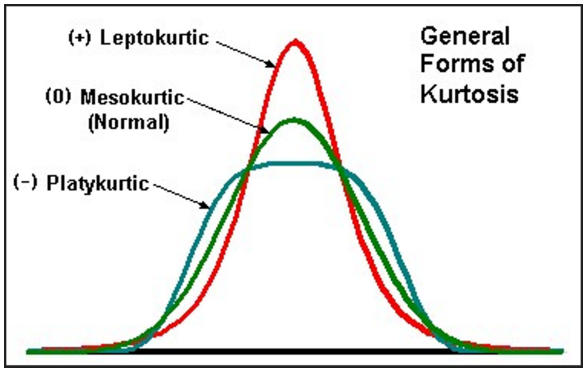
Formula for sample kurtosis:

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

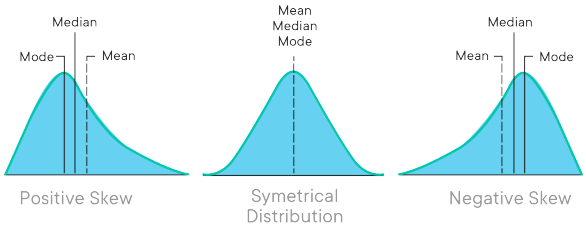
Excess kurtosis is often used:

$$\text{Excess Kurtosis} = \text{Kurtosis} - 3$$

A normal distribution has an excess kurtosis of 0.



(a) Kurtosis graph



(b) Skewness graph

Figure 1.3: Skewness and Kurtosis graphs

1.11 Estimation [24, (Pagano et all,. 2022)]

Estimation is a branch of statistical inference that deals with approximating population parameters using sample data.

1.11.1 Point Estimation

A point estimate is a single numerical value used as an approximation of an unknown parameter.

Definition

Let $X_1, X_2, \ldots, X_n$ be a random sample. A **statistic** $T(X_1, \ldots, X_n)$ is called an estimator of the parameter θ.

Examples

- Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ estimates the population mean μ
- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ estimates the population variance σ^2

Properties of a Good Estimator

Bias

The bias of an estimator $\hat{\theta}$ is defined as:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

An estimator is **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$.

Mean Squared Error (MSE)

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$$

1.11.2 Confidence Interval [20, (Rosner, 2006)]

Instead of giving a single value, an interval estimate provides a range that is likely to contain the true parameter with a certain level of confidence.

Confidence Interval for the Mean

Case 1: Variance σ^2 Known (Normal Distribution or Large Sample)

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

Where:

- $\bar{X}$ is the sample mean
- σ is the known population standard deviation
- $z_{\alpha/2}$ is the critical value from the standard normal distribution
- n is the sample size

Case 2: Variance Unknown (Small Sample from Normal Population)

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

Where:

- S is the sample standard deviation
- $t_{\alpha/2,n-1}$ is the critical value from the Student's t-distribution with $n-1$ degrees of freedom

Confidence Interval for the variance

We often want to estimate the population variance σ^2 or standard deviation σ. If the underlying population is normally distributed, confidence intervals can be derived using the chi-square distribution.

Case 1: Population Mean Unknown (use sample mean $\bar{X}$)

The statistic

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

follows a chi-square distribution with $n-1$ degrees of freedom.

Confidence Interval for Variance σ^2

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

Confidence Interval for Standard Deviation σ

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

Example

Let $n = 10$, $S^2 = 4.5$, and 95% confidence level. From chi-square tables:

$$\chi^2_{0.025,9} = 19.023, \quad \chi^2_{0.975,9} = 2.700$$

Then the confidence interval for variance:

$$\left[\frac{9 \times 4.5}{19.023},\ \frac{9 \times 4.5}{2.700}\right] = [2.13,\ 15.0]$$

For standard deviation:

$$\left[\sqrt{2.13},\ \sqrt{15.0}\right] = [1.46,\ 3.87]$$

Case 2: Population Mean Known (use true μ)

If μ is known, we use:

$$\chi^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$$

which follows a chi-square distribution with n degrees of freedom.

Confidence Interval for Variance σ^2

$$\left[\frac{\sum(X_i - \mu)^2}{\chi^2_{1-\alpha/2}},\ \frac{\sum(X_i - \mu)^2}{\chi^2_{\alpha/2}}\right]$$

Confidence Interval for Standard Deviation σ

$$\left[\sqrt{\frac{\sum(X_i - \mu)^2}{\chi^2_{1-\alpha/2}}},\ \sqrt{\frac{\sum(X_i - \mu)^2}{\chi^2_{\alpha/2}}}\right]$$

Example

Let $\mu = 10$ (known), $n = 5$, and the sample: 9, 11, 12, 8, 10.

$$\sum (X_i - \mu)^2 = (9-10)^2 + (11-10)^2 + \cdots = 10$$

From chi-square table with 5 d.o.f.:

$$\chi^2_{0.025,5} = 12.833, \quad \chi^2_{0.975,5} = 0.831$$

Then the variance interval:

$$\left[\frac{10}{12.833}, \frac{10}{0.831}\right] = [0.779,\ 12.03]$$

And standard deviation interval:

$$\left[\sqrt{0.779},\ \sqrt{12.03}\right] = [0.88,\ 3.47]$$

Confidence Interval for a Proportion p (Large n)

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

Where:

- $\hat{p}$ is the sample proportion
- n is the sample size
- The normal approximation is valid if $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$

Interpretation

A $100(1-\alpha)\%$ confidence interval means that, in repeated sampling, approximately $100(1-\alpha)\%$ of the intervals will contain the true parameter.

Example (Known Variance)

Let $\bar{X} = 50$, $\sigma = 10$, $n = 100$, and we want a 95% confidence interval:

$$z_{0.025} = 1.96, \quad \text{Margin of error} = 1.96 \cdot \frac{10}{\sqrt{100}} = 1.96$$

$$\Rightarrow [50 - 1.96,\ 50 + 1.96] = [48.04,\ 51.96]$$

Exercises

Exercise 1. ***Context.*** *Heights of wheat plants measured at heading in one plot.*
Data (cm). *52, 49, 51, 47, 55, 60, 58, 53, 52, 50, 48, 61, 57, 54, 56*
Tasks.

1. *Identify the variable type and measurement scale.*

2. *Compute: mean, median, sample variance, sample standard deviation, coefficient of variation.*

3. *Compute Q_1, Q_3, IQR, the 10th and 90th percentiles.*

4. *Using software, compute sample skewness and sample kurtosis. Interpret shape relative to a normal distribution.*

Exercise 2. ***Context.*** *Three routine measurements in agronomy.*

a) Germination test of 50 seeds from a batch with historical germination near 90 percent.

b) Insect trap counts per day with an average near 3.2 individuals.

c) Thousand-kernel weight (g) with mean near 35 and standard deviation near 4.

Tasks.

1. *For (a), model $X \sim Binomial(n = 50, p = 0.90)$. Compute $P(X \geq 45)$ and $P(40 \leq X \leq 47)$.*

2. *For (b), model $Y \sim Poisson(\lambda = 3.2)$. Compute $P(Y = 0)$ and $P(Y \geq 5)$.*

3. *For (c), model $W \sim \mathcal{N}(\mu = 35, \sigma = 4)$. Compute $P(W > 42)$ and the 5th percentile.*

Chapter 2

Statistical Test

Choosing the Right Statistical Test

Selecting an appropriate statistical test requires careful consideration of several aspects:

1. Type of Variables

Identify whether your variables are:

- **Quantitative**: numerical (continuous or discrete)
- **Qualitative**: categorical (binary, nominal with multiple categories, or ordinal)

2. Type of Measurement

Determine the specific measure you are analyzing:

- Mean
- Variance
- Proportion or percentage
- Frequency or count
- Other indicators

3. Research Objective

Clarify what comparison is needed:

- A sample versus a known population
- Two independent groups
- More than two groups

4. Relationship Between Samples

Assess whether the samples are:

- **Paired**: related or matched observations
- **Independent**: no relation between groups

5. Sample Size

Consider the size of your dataset:

- **Large sample**: often allows for parametric tests
- **Small sample**: may require non-parametric methods

6. Statistical Assumptions

Check if your data meet the conditions for test validity:

- Are the data **normally distributed**?
- Are the **variances equal** across groups?
- Is the **sample size sufficient** for the test?

Key Statistical Tests for Comparison and Their Areas of Application

Statistical comparison tests are used to examine whether there are significant differences between distributions in terms of means, variances, percentages, and other characteristics.

The core idea behind these tests is to formulate a **null hypothesis** (H_0)—typically stating that there is no difference between parameters—and decide whether to **accept** it or **reject** it in favor of an **alternative hypothesis** (H_1) based on the data.

The choice of test depends on the nature of the comparison and involves selecting the appropriate probability distribution:

1. Z-test (Standard Normal Distribution)

Applied when population variance is known or the sample size is large. Suitable for:

- Comparing two means
- Comparing paired means from two related samples
- Comparing a sample mean to a known population mean

2. t-test (Student's t-distribution)

Used when variances are unknown and/or sample sizes are small. Applicable for:

- Comparing two independent sample means
- Comparing paired observations
- Testing a sample mean against a theoretical value

3. F-test (Fisher-Snedecor Distribution)

Used to compare variances or in analysis of variance (ANOVA). Appropriate for:

- Comparing two variances
- Comparing multiple means across groups (ANOVA)
- Comparing two proportions

4. Chi-Square Test (χ^2 Distribution)

Often used with categorical data to test differences in distributions. Useful for:

- Comparing an observed frequency distribution to an expected one (goodness-of-fit)
- Comparing multiple categorical distributions (test of independence)
- Testing for differences between several proportions

2.1 Z-Test for Comparing a Sample Mean to a Known Value ($n \geq 30$) [20, (Rosner, 2006)]

The Z-test is used to determine whether the mean of a sample significantly differs from a known or hypothesized population mean.

1. Conditions

- The variable x is **quantitative**.
- The sample is **large** (typically $n \geq 30$).
- The sample has:
  - Mean: $\bar{x}$
  - Standard deviation: s

We want to test whether the observed sample mean $\bar{x}$ significantly differs from the known population mean m.

2. Hypotheses

Depending on the research question, we define:

- **Two-tailed test**:
$$H_0 : \bar{x} = m \quad \text{vs.} \quad H_1 : \bar{x} \neq m$$
- **One-tailed test**:
$$H_0 : \bar{x} = m \quad \text{vs.} \quad H_1 : \bar{x} > m \quad \text{or} \quad \bar{x} < m$$

3. Test Statistic

We compute the Z-score:

$$z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}}$$

Under the null hypothesis, z follows the standard normal distribution.

4. Critical Values

We compare the calculated Z-score to the critical values:

Test Type	Significance Level	Critical Value(s)
Two-tailed	$\alpha = 0.05$	$z_{0.025} = \pm 1.96$
One-tailed	$\alpha = 0.05$	$z_{0.05} = \pm 1.65$

5. Decision Rule

Hypothesis Type	Condition	Decision
Two-tailed	$\lvert z \rvert < z_{\alpha/2}$	Do not reject H_0
Two-tailed	$\lvert z \rvert \geq z_{\alpha/2}$	Reject H_0
One-tailed	$z < z_\alpha$ or $>$	Do not reject H_0
One-tailed	$z \geq z_\alpha$ or $<$	Reject H_0

Note

Critical values z_α and $z_{\alpha/2}$ are obtained from the standard normal distribution (Z-table).

2.2 t-Test for Comparing a Sample Mean to a Known Value (Small Sample Size) [18, (Norman and Streiner, 2008)]

Objective

The t-test is used to determine whether the mean of a small sample significantly differs from a known or hypothesized population mean.

1. Conditions

- The variable x is **quantitative**.
- The sample is **small** (typically $n < 30$).
- The population standard deviation is **unknown**.
- The sample has:
  - Mean: $\bar{x}$
  - Sample standard deviation: s
- The population is assumed to follow a **normal distribution**.

We want to test whether the observed sample mean $\bar{x}$ significantly differs from the known population mean m.

2. Hypotheses

Depending on the research question, we define:

- **Two-tailed test**:
$$H_0 : \bar{x} = m \quad \text{vs.} \quad H_1 : \bar{x} \neq m$$
- **One-tailed test**:
$$H_0 : \bar{x} = m \quad \text{vs.} \quad H_1 : \bar{x} > m \quad \text{or} \quad \bar{x} < m$$

3. Test Statistic

We compute the t-statistic:
$$t = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}}$$

Under the null hypothesis, t follows the **Student's t-distribution** with $n - 1$ degrees of freedom.

4. Critical Values

We compare the calculated t-value to the critical values from the t-distribution table, based on the significance level and degrees of freedom ($df = n - 1$):

Test Type	Significance Level	Critical Value(s)
Two-tailed	$\alpha = 0.05$	$t_{0.025,\, n-1}$
One-tailed	$\alpha = 0.05$	$t_{0.05,\, n-1}$

5. Decision Rule

Hypothesis Type	Condition	Decision
Two-tailed	$\|t\| < t_{\alpha/2,n-1}$	Do not reject H_0
Two-tailed	$\|t\| \geq t_{\alpha/2,n-1}$	Reject H_0
One-tailed	$t < t_{\alpha,n-1}$ or $>$	Do not reject H_0
One-tailed	$t \geq t_{\alpha,n-1}$ or $<$	Reject H_0

Note

Critical values t_α and $t_{\alpha/2}$ are obtained from the Student's t-distribution table based on degrees of freedom $df = n-1$. The t-test assumes the population distribution is approximately normal.

2.3 Z-Test for Comparing Two Independent Means (Large Samples) [18, (Norman and Streiner, 2008)]

Objective

To determine whether there is a statistically significant difference between the means of two independent groups using a Z-test, when both sample sizes are large.

1. Conditions

- Two independent random samples are drawn from two populations.
- Sample sizes: $n_1 \geq 30$ and $n_2 \geq 30$.
- Sample means: $\bar{x}_1$, $\bar{x}_2$.
- Sample standard deviations: s_1, s_2.
- We assume population variances are unknown but approximated by sample variances.

2. Hypotheses

- **Null hypothesis** (H_0): $\mu_1 = \mu_2$
- **Two-tailed alternative** (H_1): $\mu_1 \neq \mu_2$
- **One-tailed alternative** (H_1): $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

3. Test Statistic

We compute the Z-score as:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This test statistic follows the standard normal distribution under the null hypothesis.

4. Critical Values

We compare the computed Z-value to the critical values of the standard normal distribution, depending on the significance level α.

Test Type	Significance Level	Critical Value(s)
Two-tailed	$\alpha = 0.05$	$z_{0.025} = \pm 1.96$
One-tailed	$\alpha = 0.05$	$z_{0.05} = \pm 1.65$

5. Decision Rule

Hypothesis Type	Condition	Decision
Two-tailed	$\lvert z \rvert < z_{\alpha/2}$	Do not reject H_0
Two-tailed	$\lvert z \rvert \geq z_{\alpha/2}$	Reject H_0
One-tailed	$z < z_\alpha$ or $>$	Do not reject H_0
One-tailed	$z \geq z_\alpha$ or $<$	Reject H_0

Note

The critical values z_α and $z_{\alpha/2}$ are found in the standard normal (Z) table. This Z-test is valid when sample sizes are large and the central limit theorem applies.

2.4 t-Test for Two Independent Samples (Small Sample Sizes) [18, (Norman and Streiner, 2008)]

Objective

To determine whether the means of two independent populations are significantly different when both sample sizes are small ($n_1 < 30$, $n_2 < 30$), using the Student's t-test.

1. Conditions

- Two independent samples are drawn: sample 1 of size n_1, and sample 2 of size n_2, both with $n < 30$.

- Sample means: $\bar{x}_1$, $\bar{x}_2$
- Sample standard deviations: s_1, s_2
- Equal population variances are assumed.

2. Hypotheses

- **Null hypothesis** (H_0): $\mu_1 = \mu_2$
- **Two-tailed alternative** (H_1): $\mu_1 \neq \mu_2$
- **One-tailed alternative** (H_1): $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

3. Test Statistic

We compute the t-statistic as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The degrees of freedom are:

$$\text{df} = n_1 + n_2 - 2$$

4. Decision Rule

Compare the calculated t-value with the critical value from the Student's t-distribution for the chosen significance level α and degrees of freedom df.

Hypothesis Type	Condition	Decision
Two-tailed	$\lvert t\rvert < t_{\alpha/2,\,df}$	Do not reject H_0
Two-tailed	$\lvert t\rvert \geq t_{\alpha/2,\,df}$	Reject H_0
One-tailed	$t < t_{\alpha,\,df}$ or $>$	Do not reject H_0
One-tailed	$t \geq t_{\alpha,\,df}$ or $<$	Reject H_0

5. Note

Critical values t_α and $t_{\alpha/2}$ are obtained from the t-distribution table based on degrees of freedom. This test assumes normal distribution of data and equal population variances.

2.5 Z-Test for Comparing Two Paired Means (related paired) [18, (Norman and Streiner, 2008)]

Objective

To determine whether there is a statistically significant difference between the means of two related (paired) measurements observed on the same individuals.

1. Conditions

- A single sample of size $n \geq 30$.
- Two related (paired) measurements are recorded for each subject: x_i and y_i, for $i = 1, \ldots, n$.
- For each pair, the difference $d_i = x_i - y_i$ is calculated.
- We analyze the mean $\bar{d}$ and the standard deviation s_d of the differences.

2. Hypotheses

- **Null hypothesis** (H_0): $\bar{d} = 0$ (no difference in means)
- **Two-tailed alternative** (H_1): $\bar{d} \neq 0$
- **One-tailed alternative** (H_1): $\bar{d} > 0$ or $\bar{d} < 0$

3. Test Statistic

We compute the Z-score for the mean of the paired differences:

$$z = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

where:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad s_d^2 = \frac{1}{n-1}\left(\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right)$$

4. Critical Values

The calculated Z-value is compared to the critical values of the standard normal distribution.

Test Type	**Significance Level**	**Critical Value(s)**
Two-tailed	$\alpha = 0.05$	$z_{0.025} = \pm 1.96$
One-tailed	$\alpha = 0.05$	$z_{0.05} = \pm 1.65$

5. Decision Rule

Hypothesis Type	**Condition**	**Decision**
Two-tailed	$\lvert z \rvert < z_{\alpha/2}$	Do not reject H_0: no significant difference
Two-tailed	$\lvert z \rvert \geq z_{\alpha/2}$	Reject H_0: significant difference
One-tailed	$z < z_\alpha$ or $>$	Do not reject H_0
One-tailed	$z \geq z_\alpha$ or $<$	Reject H_0: one mean is significantly greater or smaller

Note

This Z-test for paired samples is valid when the number of observations is large ($n \geq 30$) and the differences between pairs are approximately normally distributed. For smaller sample sizes, a t-test should be used.

2.6 Chi-Square Test for a Single Variance (Large Sample) [18, (Norman and Streiner, 2008)]

Objective

To test whether the variance of a population is equal to a specified theoretical value when the sample size is large ($n > 30$), using the chi-square distribution.

1. Conditions

- A single random sample of size $n > 30$.
- The variable is quantitative.
- The sample variance is s^2, and the hypothesized population variance is σ_0^2.
- The population is approximately normal (less critical due to large n).

2. Hypotheses

- **Null hypothesis** (H_0): $\sigma^2 = \sigma_0^2$
- **Two-tailed alternative** (H_1): $\sigma^2 \neq \sigma_0^2$
- **One-tailed alternative** (H_1): $\sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$

3. Test Statistic

$$\chi^2 = \frac{(n-1) \cdot s^2}{\sigma_0^2}$$

Under the null hypothesis, this statistic follows a chi-square distribution with $df = n - 1$ degrees of freedom.

4. Decision Rule

Use the chi-square distribution table to find the critical values at the significance level α, with $df = n - 1$.

Hypothesis Type	Condition	Decision
Two-tailed	$\chi^2_{\alpha/2} < \chi^2 < \chi^2_{1-\alpha/2}$	Do not reject H_0
Two-tailed	$\chi^2 \leq \chi^2_{\alpha/2}$ or $\chi^2 \geq \chi^2_{1-\alpha/2}$	Reject H_0
One-tailed (right)	$\chi^2 < \chi^2_{1-\alpha}$	Do not reject H_0
One-tailed (right)	$\chi^2 \geq \chi^2_{1-\alpha}$	Reject H_0
One-tailed (left)	$\chi^2 > \chi^2_{\alpha}$	Do not reject H_0
One-tailed (left)	$\chi^2 \leq \chi^2_{\alpha}$	Reject H_0

5. Notes

- The test is more robust for large samples, even if the normality assumption is not strictly met.
- Always check the critical values from chi-square distribution tables with $df = n - 1$.

2.7 Chi-Square Test for a Single Variance (Small Sample) [18, (Norman and Streiner, 2008)]

Objective

To test whether the variance of a population is equal to a specified value when the sample size is small ($n < 30$) using the chi-square distribution.

1. Conditions

- A single sample of size $n < 30$.
- The variable is quantitative.
- The population from which the sample is drawn is normally distributed.
- The sample variance is s^2, and the hypothesized population variance is σ_0^2.

2. Hypotheses

- **Null hypothesis** (H_0): $\sigma^2 = \sigma_0^2$
- **Two-tailed alternative** (H_1): $\sigma^2 \neq \sigma_0^2$
- **One-tailed alternative** (H_1): $\sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$

3. Test Statistic

The chi-square test statistic is given by:

$$\chi^2 = \frac{(n-1) \cdot s^2}{\sigma_0^2}$$

It follows a chi-square distribution with $df = n - 1$ degrees of freedom under the null hypothesis.

4. Decision Rule

Compare the computed chi-square value to the critical value(s) from the chi-square distribution table for the corresponding degrees of freedom and significance level.

Hypothesis Type	**Condition**	**Decision**
Two-tailed	$\chi^2_{\alpha/2} < \chi^2 < \chi^2_{1-\alpha/2}$	Do not reject H_0
Two-tailed	$\chi^2 \leq \chi^2_{\alpha/2}$ or $\chi^2 \geq \chi^2_{1-\alpha/2}$	Reject H_0
One-tailed (right)	$\chi^2 < \chi^2_{1-\alpha}$	Do not reject H_0
One-tailed (right)	$\chi^2 \geq \chi^2_{1-\alpha}$	Reject H_0
One-tailed (left)	$\chi^2 > \chi^2_{\alpha}$	Do not reject H_0
One-tailed (left)	$\chi^2 \leq \chi^2_{\alpha}$	Reject H_0

5. Notes

- The test is valid only if the underlying population is normal.
- For small samples, normality is especially critical.
- Critical values depend on the degrees of freedom $df = n - 1$, and are obtained from chi-square tables.

2.8 F-Test for Comparing Two Variances [18, (Norman and Streiner, 2008)]

Objective

The F-test is used to determine whether two independent populations have significantly different variances.

1. Conditions

- Two independent samples of sizes n_1 and n_2, regardless of their size.
- The data in each sample are quantitative and normally distributed.
- The sample variances are denoted by s_1^2 and s_2^2.

2. Hypotheses

Let σ_1^2 and σ_2^2 be the population variances.

- **Null hypothesis** (H_0): $\sigma_1^2 = \sigma_2^2$

- **Two-tailed alternative** (H_1): $\sigma_1^2 \neq \sigma_2^2$
- **One-tailed alternative** (H_1): $\sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$

3. Test Statistic

We define the F-statistic as:

$$F = \frac{s^2_{\text{larger}}}{s^2_{\text{smaller}}}$$

where: - s^2_{larger} is the larger of the two sample variances to ensure $F \geq 1$, - Degrees of freedom: $\text{df}_1 = n_1 - 1$, $\text{df}_2 = n_2 - 1$

4. Decision Rule

Using the F-distribution table, we compare the calculated value with the critical value(s) for the chosen significance level α.

Hypothesis Type	**Condition**	**Decision**
Two-tailed	$F < F_{1-\alpha/2,df_1,df_2}$	Do not reject H_0
Two-tailed	$F \geq F_{1-\alpha/2,df_1,df_2}$	Reject H_0
One-tailed	$F < F_{\alpha,df_1,df_2}$	Do not reject H_0
One-tailed	$F \geq F_{\alpha,df_1,df_2}$	Reject H_0

5. Notes

- The F-distribution is not symmetric. Always place the larger variance in the numerator.
- Use the F-table to find the critical value based on α, df_1, and df_2.
- The test assumes normality in both populations.

2.9 Z-Test for a Single Proportion ($n > 30$) [5, (Gerstman, 2014)]

Objective

To test whether the observed proportion in a large sample differs significantly from a known or hypothesized population proportion.

1. Conditions

- A single large random sample ($n > 30$).
- A binary variable (success/failure).

- Let:
  - x: number of successes in the sample.
  - n: sample size.
  - $\hat{p} = \frac{x}{n}$: observed sample proportion.
  - p_0: hypothesized population proportion.
- The sample satisfies the normal approximation conditions:

$$np_0 \geq 5 \quad \text{and} \quad n(1-p_0) \geq 5$$

2. Hypotheses

- **Null hypothesis** (H_0): $p = p_0$
- **Two-tailed alternative** (H_1): $p \neq p_0$
- **One-tailed alternative** (H_1): $p > p_0$ or $p < p_0$

3. Test Statistic

The Z-score is computed as:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Note: Under the null hypothesis, z follows the standard normal distribution.

4. Decision Rule

Use the critical values from the standard normal (Z) distribution to make your decision.

Hypothesis Type	**Condition**	**Decision**
Two-tailed	$\|z\| < z_{\alpha/2}$	Do not reject H_0
Two-tailed	$\|z\| \geq z_{\alpha/2}$	Reject H_0
One-tailed	$z < z_\alpha$ or $z > -z_\alpha$	Do not reject H_0
One-tailed	$z \geq z_\alpha$ or $z \leq -z_\alpha$	Reject H_0

5. Notes

- The Z-test is reliable for large samples due to the central limit theorem.
- Common critical values:

$$z_{0.025} = 1.96 \quad \text{(two-tailed)}, \quad z_{0.05} = 1.65 \quad \text{(one-tailed)}$$

- If conditions are not met (e.g., small n), consider using the binomial test.

2.10 Z-Test for a Single Proportion ($n < 30$) [5, (Gerstman, 2014)]

Objective

To test whether the observed proportion in a small sample differs significantly from a known or hypothesized population proportion.

1. Conditions

- A single sample of size $n < 30$.
- A binary variable (success/failure).
- Let:
  - x: number of observed successes.
  - n: sample size.
  - $\hat{p} = \frac{x}{n}$: sample proportion.
  - p_0: hypothesized population proportion.
- Caution: The Z-test may be inaccurate for small n — use exact binomial test if necessary.

2. Hypotheses

- **Null hypothesis** (H_0): $p = p_0$
- **Two-tailed alternative** (H_1): $p \neq p_0$
- **One-tailed alternative** (H_1): $p > p_0$ or $p < p_0$

3. Test Statistic (Z-approximation)

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Note: This formula assumes that the sample size is still large enough to approximate the binomial distribution with a normal one, i.e., $np_0 \geq 5$ and $n(1 - p_0) \geq 5$. If not, use the binomial test.

4. Decision Rule

Compare the test statistic z to the critical values from the standard normal distribution.

Hypothesis Type	Condition	Decision
Two-tailed	$\|z\| < z_{\alpha/2}$	Do not reject H_0
Two-tailed	$\|z\| \geq z_{\alpha/2}$	Reject H_0
One-tailed	$z < z_\alpha$ or $z > -z_\alpha$	Do not reject H_0
One-tailed	$z \geq z_\alpha$ or $z \leq -z_\alpha$	Reject H_0

5. Notes

- For $n < 30$, prefer an **exact binomial test** if normality conditions are not met.
- The Z-test is acceptable only if $np_0 \geq 5$ and $n(1 - p_0) \geq 5$.
- Common critical values: $z_{0.025} = 1.96$, $z_{0.05} = 1.65$.

2.11 Z-Test for Comparing Two Proportions [5, (Gerstman, 2014)]

Objective

To test whether the proportions of a characteristic are significantly different in two independent populations.

1. Conditions

- Two independent random samples.
- Each sample includes a binary variable (success/failure).
- Let:
  - n_1, n_2: sample sizes
  - x_1, x_2: number of successes in each sample
  - $p_1 = \frac{x_1}{n_1}$, $p_2 = \frac{x_2}{n_2}$: sample proportions
- Sample sizes should be large enough to assume normal approximation: $n_1p_1, n_1(1 - p_1), n_2p_2, n_2(1 - p_2) \geq 5$

2. Hypotheses

- **Null hypothesis** (H_0): $p_1 = p_2$
- **Two-tailed alternative** (H_1): $p_1 \neq p_2$
- **One-tailed alternative** (H_1): $p_1 > p_2$ or $p_1 < p_2$

3. Test Statistic

- First, compute the pooled proportion:

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

- Then compute the Z-score:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

4. Decision Rule

Compare the calculated Z-score with critical values from the standard normal distribution (Z-distribution).

Hypothesis Type	Condition	Decision
Two-tailed	$\|z\| < z_{\alpha/2}$	Do not reject H_0
Two-tailed	$\|z\| \geq z_{\alpha/2}$	Reject H_0
One-tailed	$z < z_\alpha$ or $>$	Do not reject H_0
One-tailed	$z \geq z_\alpha$ or $<$	Reject H_0

5. Notes

- The test is valid only if the sample sizes are sufficiently large for the normal approximation.
- Critical values for z at the 5% level: $z_{0.025} = \pm 1.96$, $z_{0.05} = \pm 1.65$
- The method assumes independent sampling.

2.12 Chi-square Tests [25, (Williams, 2017)]

2.12.1 Goodness-of-Fit Test (One Distribution vs. Theoretical)

Objective:

To test whether an observed categorical distribution differs significantly from a theoretical (expected) distribution.

1. Conditions:

- The data represent observed frequencies across k categories.
- The expected frequencies are calculated using known probabilities P_i such that $\sum P_i = 1$.
- All expected frequencies must be ≥ 5.

2. Hypotheses:

- H_0: The observed distribution matches the expected distribution.
- H_1: The observed distribution does not match the expected distribution.

3. Test Statistic:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(f_{oi} - f_{ti})^2}{f_{ti}}$$

Where:

- f_{oi}: observed frequency in category i
- $f_{ti} = n \cdot P_i$: expected frequency for category i

4. Decision Rule:

- Degrees of freedom: $k - 1$
- Compare χ_c^2 to the critical value χ_α^2 from the chi-square table.
- If $\chi_c^2 \geq \chi_\alpha^2$, reject H_0 (significant difference).

2.12.2 Chi-square Test of Homogeneity (Comparing Distributions Between Samples)

Objective:

To compare categorical distributions (proportions) across multiple independent groups.

1. Conditions:

- Observed data are arranged in a two-dimensional contingency table.
- All expected counts must be at least 5.

2. Hypotheses:

- H_0: The distributions (proportions) are the same across groups.
- H_1: At least one group differs significantly in distribution.

3. Test Statistic:

$$\chi_c^2 = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(f_{oij} - f_{tij})^2}{f_{tij}}$$

Where:

- f_{oij}: observed frequency in category i, group j
- f_{tij}: expected frequency, usually calculated as:

$$f_{tij} = \frac{(\text{row total}_i) \cdot (\text{column total}_j)}{\text{grand total}}$$

4. Decision Rule:

- Degrees of freedom: $(k-1)(c-1)$
- Compare χ_c^2 to χ_α^2.
- If $\chi_c^2 \geq \chi_\alpha^2$, reject H_0 (significant difference in proportions).

Exercises

Exercise 3. ***Context.*** *Soil nitrate in mg/kg measured after preplant fertilization. Agronomic guideline: $\mu_0 = 25$ mg/kg.*
Data. *28, 24, 27, 31, 26, 29, 22, 30, 25, 27, 26, 28, 24, 33, 29*
Tasks.

1. *Test $H_0 : \mu = 25$ against $H_1 : \mu \neq 25$ at $\alpha = 0.05$ using a one-sample t test. Report t, df, p-value, and a 95% confidence interval.*

Exercise 4. ***Context.*** *Disease incidence in two tomato cultivars under the same management.*
Data.
Cultivar A: $n = 120$ plants, infected $= 18$
Cultivar B: $n = 110$ plants, infected $= 33$
Tasks.

1. *Test $H_0 : p_A = p_B$ using a large-sample z test for two proportions. Report the estimated difference, its 95% confidence interval, the test statistic, and p-value.*

Chapter 3

Comparison of several means (Fisher's F-Test)

3.1 One-Way ANOVA Test [23, (Sokal, R. R and Rohlf, F. J, 1987)]

1. Objective and Conditions

The Fisher F-test (ANOVA) is used to compare the means of a quantitative variable across multiple independent samples.

Assumptions:

- Each sample comes from a normally distributed population.
- Populations have equal variances (homoscedasticity).
- Samples are independent.

2. Hypotheses

Let c be the number of groups and μ_i the mean of the i^{th} population.

- Null hypothesis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_c$
- Alternative hypothesis: H_1 : At least one mean is different

3. Computation

We calculate the ratio of the variance **between groups** to the variance **within groups**:

$$F = \frac{s_g^2}{s_r^2}$$

Where:

$$s_g^2 = \frac{\sum \frac{x_i^2}{n_i} - \frac{x_g^2}{N}}{c-1} \qquad \text{and} \qquad s_r^2 = \frac{\sum x^2 - \sum \frac{x_i^2}{n_i}}{N-c}$$

Notation:

- x_i: Sum of observed values in group i
- n_i: Sample size of group i
- x_g: Total sum of all observed values
- N: Total sample size from all groups
- c: Number of groups
- $\sum x^2$: Sum of squared individual observations

Degrees of freedom:

$$\mathrm{df}_1 = c - 1, \quad \mathrm{df}_2 = N - c$$

4. Decision Rule

Using the F-distribution table at significance level $\alpha = 5\%$:

- If $F < F_{\alpha, c-1, N-c} \rightarrow$ do not reject H_0: means are not significantly different
- If $F \geq F_{\alpha, c-1, N-c} \rightarrow$ reject H_0: at least one mean differs significantly

5. Conclusion

The Fisher's F-test (ANOVA) is a robust method for determining if observed differences between group means are statistically significant, under the assumption of normality and homogeneity of variance.

3.2 Two-Way ANOVA Test [23, (Sokal, R. R and Rohlf, F. J, 1987)]

1. Objective

Two-way ANOVA is used to study the effects of two independent categorical factors on a quantitative dependent variable, and to determine if there is an interaction effect between the two factors.

2. Experimental Design and Conditions

- The dependent variable is quantitative and normally distributed within each group.
- The independent variables (factors) are qualitative:
  - Factor A has a levels.
  - Factor B has b levels.
- The observations are independent.
- Homogeneity of variances (homoscedasticity) is assumed.
- The design may be balanced (equal group sizes) or unbalanced.

3. Hypotheses

We test the following:

- Main effect of Factor A:

$$H_0^A : \text{All means of Factor A are equal} \quad \text{vs.} \quad H_1^A : \text{At least one mean differs}$$

- Main effect of Factor B:

$$H_0^B : \text{All means of Factor B are equal} \quad \text{vs.} \quad H_1^B : \text{At least one mean differs}$$

- Interaction effect:

$$H_0^{AB} : \text{No interaction between A and B} \quad \text{vs.} \quad H_1^{AB} : \text{Interaction exists}$$

4. ANOVA Table Structure

Source of Variation	**Sum of Squares (SS)**	**Degrees of Freedom (df)**	**Mean Square (MS)**
Factor A	SSA	$a-1$	$\text{MSA} = \frac{\text{SSA}}{a-1}$
Factor B	SSB	$b-1$	$\text{MSB} = \frac{\text{SSB}}{b-1}$
Interaction (A×B)	SSAB	$(a-1)(b-1)$	$\text{MSAB} = \frac{\text{SSAB}}{(a-1)(b-1)}$
Error (Residual)	SSE	$ab(n-1)$	$\text{MSE} = \frac{\text{SSE}}{ab(n-1)}$
Total	SST	$N-1$	–

Table 3.1: Two-way ANOVA decomposition of variance

5. Test Statistics

$$F_A = \frac{\text{MSA}}{\text{MSE}}, \quad F_B = \frac{\text{MSB}}{\text{MSE}}, \quad F_{AB} = \frac{\text{MSAB}}{\text{MSE}}$$

These F-values are compared to critical values from the F-distribution with the corresponding degrees of freedom and significance level α (typically 0.05).

6. Decision Rules

- If $F > F_{\text{crit}}$, reject the null hypothesis for that source of variation.
- Otherwise, do not reject the null hypothesis.

7. Interpretation

- If the interaction is significant, interpret the interaction plot and ignore the main effects.
- If the interaction is not significant, interpret the main effects independently.

8. Notes

- For balanced designs (equal observations per group), calculations are simplified.
- Post-hoc tests (like Tukey HSD) are used if main effects are significant and the number of levels > 2.

Key Formula Summary

Total Sum of Squares (SST)

$$\text{SST} = \sum_{i,j,k} (X_{ijk} - \bar{X})^2$$

ANOVA Decomposition

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}$$

- **SSA**: Sum of squares due to Factor A (main effect)
- **SSB**: Sum of squares due to Factor B (main effect)
- **SSAB**: Sum of squares due to interaction between A and B
- **SSE**: Sum of squares due to error (residuals)

Formulas for Components

$$\text{SSA} = \sum_{i=1}^{a} n_b (\bar{X}_{i\cdot\cdot} - \bar{X})^2$$

$$\text{SSB} = \sum_{j=1}^{b} n_a (\bar{X}_{\cdot j\cdot} - \bar{X})^2$$

$$\text{SSAB} = \sum_{i=1}^{a} \sum_{j=1}^{b} n (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2$$

$$\text{SSE} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (X_{ijk} - \bar{X}_{ij\cdot})^2$$

Exercises

Exercise 5. ***Context.*** *Grain yield (t/ha) under three fertilizer programs with four replicated plots per program.*
Data (t/ha).
Program A: 4.8, 5.1, 4.9, 5.0
Program B: 5.3, 5.5, 5.4, 5.6
Program C: 5.0, 5.2, 5.1, 5.1
Tasks.

1. *Fit a one-way ANOVA. Report the ANOVA table, F statistic, df, and p-value.*

Exercise 6. ***Context.*** *Effect of irrigation regime and cultivar on yield (t/ha). Two replicates per cell.*
Data (t/ha).

- *Irrigation: Deficit C1: 4.5, 4.7; C2: 4.8, 4.6; C3: 4.4, 4.5*
- *Irrigation: Full C1: 5.2, 5.1; C2: 5.6, 5.5; C3: 5.0, 5.1*

Tasks.

1. *Fit a two-way ANOVA with interaction. Report F tests for main effects and interaction.*
2. *Interpret the interaction plot. Does the effect of irrigation depend on cultivar?*

Chapter 4

Bivariate analysis

Definition 1. *Bivariate analysis is the statistical analysis of two variables to understand the relationship or association between them.*

4.1 Linear Regression, Covariance, and Correlation [4, (Gaddis, M. L and Gaddis, G. M, 1990)]

1. Objective

Linear regression is a statistical method used to model the relationship between a dependent variable y and one or more independent variables x. The simplest form is simple linear regression, involving a single explanatory variable.

2. Model Definition

The linear regression model is expressed as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y: dependent (response) variable
- x: independent (explanatory) variable
- β_0: intercept
- β_1: slope (effect of x on y)
- ε: random error, assumed to follow $\mathcal{N}(0, \sigma^2)$

3. Estimation of Coefficients

Using the method of least squares, the estimators for β_0 and β_1 are:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

4. Goodness of Fit

The total variability in y can be decomposed as:

$$\text{SST} = \text{SSR} + \text{SSE}$$

- SST (Total Sum of Squares): $\sum(y_i - \bar{y})^2$
- SSR (Regression Sum of Squares): $\sum(\hat{y}_i - \bar{y})^2$
- SSE (Error Sum of Squares): $\sum(y_i - \hat{y}_i)^2$

The coefficient of determination R^2 is defined as:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

It measures the proportion of variability in y explained by the regression model.

5. Hypothesis Testing

To test if x significantly affects y, we use:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

The test statistic is:

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$$

Where $\text{SE}(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$, and the test follows a Student's t-distribution with $n - 2$ degrees of freedom.

6. Assumptions

- Linearity: the relationship between x and y is linear.
- Independence: observations are independent.
- Homoscedasticity: constant variance of errors.
- Normality: errors follow a normal distribution.

4.1.1 Covariance and Correlation

The **covariance** between two variables x and y measures the joint variability:

$$\text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

To standardize this measure, we use the **Pearson correlation coefficient** r:

$$r = \frac{\text{Cov}(x, y)}{s_x s_y}$$

Where:

- s_x and s_y are the sample standard deviations of x and y
- $-1 \leq r \leq 1$

4.1.2 Linear Regression Coefficients

In the simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The estimated slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ are given by:

$$\hat{\beta}_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = r \cdot \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Interpretation:

- The slope $\hat{\beta}_1$ increases with the strength of correlation.
- The correlation coefficient r is the normalized slope when both x and y are standardized.

4.1.3 Goodness of Fit

The squared correlation coefficient r^2 is equal to the coefficient of determination in simple linear regression:

$$R^2 = r^2$$

It represents the proportion of variance in y explained by the model.

Summary Table

Quantity	**Formula**
Covariance	$\text{Cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$
Correlation	$r = \frac{\text{Cov}(x, y)}{s_x s_y}$
Slope of regression line	$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}$
Intercept	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Goodness of fit	$R^2 = r^2$

Interpretation of Covariance and Correlation

- If $\text{Cov}(x, y) > 0$, the variables tend to increase together (positive association).
- If $\text{Cov}(x, y) < 0$, one variable tends to increase as the other decreases (negative association).
- If $\text{Cov}(x, y) = 0$, the variables are linearly uncorrelated.
- $r > 0$: positive linear relationship.

- $r < 0$: negative linear relationship.
- $r = 0$: no linear relationship.
- $r \approx |1|$: perfect linear relationship.

The value of r is dimensionless and always lies between -1 and 1, which makes it easier to interpret and compare across datasets.

4.2 Scatter Plot with Regression Line

A **scatter plot** with a regression line is a visual representation of the relationship between two continuous variables. It shows:

- Individual data points (x_i, y_i)
- A regression line representing the best linear fit to the data

The regression line is computed using the method of least squares and has the form:

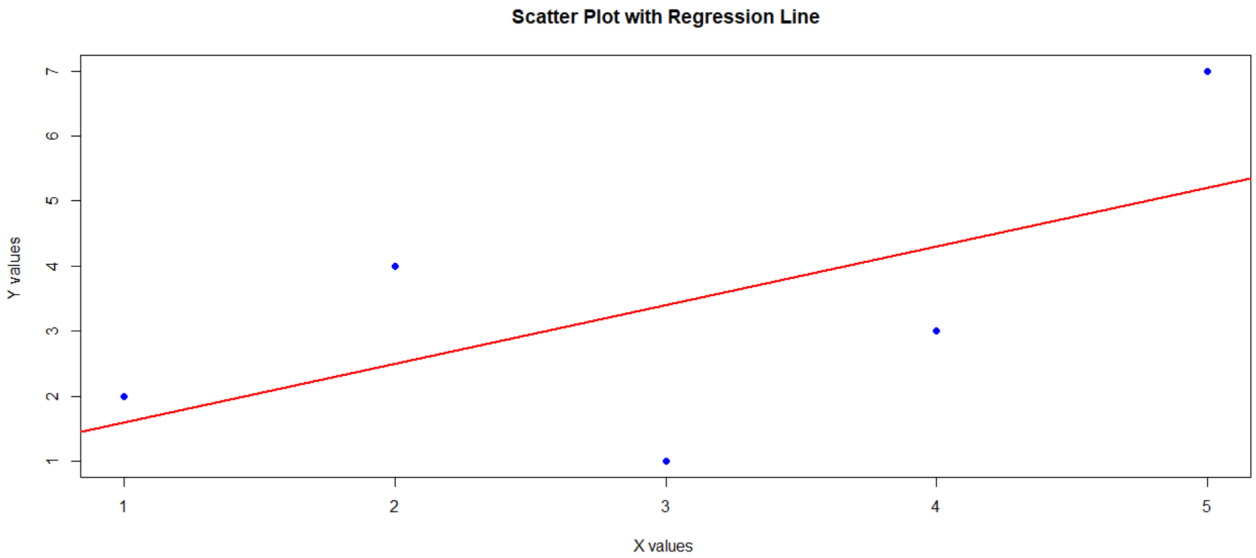
$$y = a + bx$$

where:

- a: intercept
- b: slope of the line

```r
# Example R code to create a scatter plot
x <- c(1, 2, 3, 4, 5)
y <- c(2, 4, 1, 3, 7)
plot(x, y,
     main = "Scatter Plot",
     xlab = "Variable X",
     ylab = "Variable Y",
     pch = 19,
     col = "blue")
```

Listing 4.1: R code to generate a scatter from (xi, ni)



4.3 Multiple Linear Regression [11, (Jobson, J. D, 1991)]

1. Objective

Multiple linear regression models the relationship between a response variable y and several explanatory (independent) variables $x_1, x_2, \ldots, x_p$. The aim is to estimate the contribution of each predictor to the response.

2. Model Formulation

The general form of the multiple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad \text{for } i = 1, 2, \ldots, n$$

Or, in matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Where:

- $\mathbf{y}$ is the $n \times 1$ vector of observed responses.
- $\mathbf{X}$ is the $n \times (p+1)$ matrix of predictors (including a column of 1s for the intercept).
- $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of regression coefficients.
- $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of random errors.

3. Estimation of Parameters

The regression coefficients are estimated using the 'Ordinary Least Squares (OLS)' method:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

This minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

4. Interpretation

Each coefficient β_j represents the 'expected change in the response y' for a one-unit increase in x_j, holding all other variables constant.

5. Assumptions

Multiple linear regression relies on several key assumptions:

- Linearity: The relationship between predictors and the response is linear.
- Independence: The residuals are independent.
- Homoscedasticity: The residuals have constant variance.
- Normality: The residuals are normally distributed.
- No multicollinearity: Predictors are not highly correlated.

6. Goodness of Fit

The 'coefficient of determination' R^2 measures how well the model explains the variability in y:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

- SS_{res} = Residual Sum of Squares
- SS_{tot} = Total Sum of Squares

7. Hypothesis Testing and Decision

A. Global Significance of the Model (F-test)

We test whether at least one predictor has a significant effect on the response variable.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{(no effect)}$$
$$H_1 : \text{At least one } \beta_j \neq 0$$

We use the "F-statistic":

$$F = \frac{MS_{\text{reg}}}{MS_{\text{res}}} = \frac{(SS_{\text{tot}} - SS_{\text{res}})/p}{SS_{\text{res}}/(n-p-1)}$$

Compare this value to the critical value $F_{\alpha,p,n-p-1}$ from the F-distribution table.

- If $F \geq F_\alpha$: reject H_0, the model is globally significant.
- If $F < F_\alpha$: do not reject H_0, the model is not significant.

B. Significance of Individual Predictors (t-test)

To test if an individual predictor x_j significantly contributes to the model:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

The test statistic is:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Where $SE(\hat{\beta}_j)$ is the standard error of the coefficient.
Compare $|t_j|$ to the critical value $t_{\alpha/2,n-p-1}$:

- If $|t_j| \geq t_{\alpha/2}$: reject H_0, the variable x_j is significant.
- If $|t_j| < t_{\alpha/2}$: do not reject H_0, the variable x_j is not significant.

8. Confidence Intervals for Coefficients

Each coefficient β_j has a confidence interval:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p-1} \cdot SE(\hat{\beta}_j)$$

This interval gives the range of plausible values for β_j at a given confidence level (e.g., 95%).

Exercises

Exercise 7. ***Context.*** *Relationship between growing degree days (GDD) and maize yield.* ***Data.*** *GDD: 950, 980, 1010, 1040, 1070, 1100, 1130, 1160, 1190, 1220 Yield (t/ha): 6.2, 6.5, 6.7, 6.8, 7.0, 7.2, 7.3, 7.5, 7.6, 7.7*

Tasks.

1. *Fit* $Yield = \beta_0 + \beta_1 GDD$. *Report* $\hat{\beta}_0$, $\hat{\beta}_1$, *standard errors, t tests, and* R^2.
2. *Predict yield at* $GDD = 1080$ *with a* 95% *prediction interval for a future field.*

Exercise 8. ***Context.*** *Winter wheat yield as a function of nitrogen, in-season rainfall, and plant density. Twelve fields observed.*

Data.

- *Yield (t/ha): 5.1, 5.6, 5.9, 6.2, 5.4, 6.5, 6.1, 6.8, 5.7, 6.3, 6.0, 6.6*
- *Nitrogen (kg/ha): 80, 100, 110, 130, 90, 150, 120, 160, 95, 140, 115, 155*
- *Rainfall (mm): 180, 210, 220, 240, 190, 260, 230, 270, 200, 250, 225, 265*
- *Density (plants/m^2): 220, 240, 250, 260, 230, 280, 255, 290, 235, 270, 250, 285*

Tasks.

1. *Fit* $Yield = \beta_0 + \beta_1 Nitrogen + \beta_2 Rainfall + \beta_3 Density$. *Report coefficients, standard errors, t tests, and adjusted* R^2.

Supplementary exercises

Chaptre 1

Example 1. *A researcher records the weights (in grams) of 12 tomatoes from a greenhouse crop:*

90	92	88	95	100	102	105	108	110	115	118	120

1. *Sort the data.*
2. *Find: Q1, Q2 and Q3.*
3. *Find: D1, D5 and D9.*
4. *Compute the sample mean and standard deviation.*
5. *Calculate the skewness*
6. *Interpret the skewness: Is the distribution symmetric, positively skewed, or negatively skewed?*
7. *Compute the sample kurtosis*
8. *Interpret: Is the distribution leptokurtic (sharp), platykurtic (flat), or mesokurtic (normal-like)?*

```r
# R code (Solution)
weights<-c(90,92,88,95,100,102,105,108,110,115,118,120);
sort<-sort(weights);
Q1<-summary(weights)[2];
Q2<-summary(weights)[3];
Q3<-summary(weights)[5];
deciles <- quantile(weights, probs = seq(0.1, 0.9, by = 0.1));
D1<- deciles[1];
D5<- deciles[5];
D9<- deciles[9];
Mean<-summary(weights)[4];
sd<-sd(weights);
n=length(weights);
skewness=(sum((weights-Mean)^3))/((n-1)*sd^3);
kurtosis=(sum((weights-Mean)^4))/((n-1)*sd^4);
if (skewness>0) {
 print("␣The␣distribution␣is␣right-skewed")
} else if (skewness<0) {
  print("␣The␣distribution␣is␣left-skewed")
```

```
} else {
 print("␣The␣distribution␣is␣symmetric")
}
if (kurtosis>0) {
 print("␣The␣distribution␣is␣Leptokurtic:")
} else if (kurtosis<0) {
  print("␣The␣distribution␣is␣Platykurtic␣")
} else {
 print("␣The␣distribution␣is␣Mesokurtic")
}
```

Listing 4.2: R code (Solution example 1)

Chaptre 2

Example 2. *Traditionally, kharoub syrup contains about* 54 *grams of sugar per* 100 *ml. A food technologist analyzes* 8 *batches of syrup after reformulation. The sugar contents (in g/*100*ml) are:*

52.5	53.0	54.3	51.8	52.9	53.1	52.2	53.4

- *At the* 1% *significance level, is there evidence that the mean sugar content is now equal to* 54 *grams?*

```
# R code (Solution)
kharoub<-c(52.5,53.0,54.3,51.8,52.9,53.1,52.2,53.4);
m<-54;
n<-length(kharoub);
t<-(mean(kharoub)-m)/(sd(kharoub)/sqrt(n))
alpha <- 0.05;
df <- n;
t_crit <- qt(1 - alpha/2, df);
if(abs(t)<t_crit)
{
sprintf("Do␣not␣reject␣H0␣␣␣μ= %g", m)
}
if(abs(t)>=t_crit)
{
 sprintf("Reject␣H0␣μ ≠ %g", m)
}
```

Listing 4.3: R code (Solution example 2)

Example 3. *A researcher wants to compare the effects of two fertilizers (A and B) on wheat yield.*

Fertilizer A:	3.1	2.9	3.5	3.0	3.3
Fertilizer B:	3.8	4.0	3.9	4.1	4.2

1. *Calculate the* ***mean*** *and* ***standard deviation*** *of each group.*

2. *Perform a* ***two-sample t-test*** *at the 5% significance level.*

3. *Interpret the result in the context of fertilizer efficiency.*

```r
# R code (Solution)
FertilizerA<-c(3.1,2.9,3.5,3.0,3.3);
FertilizerB<-c(3.8,4.0,3.9,4.1,4.2);
x1<-mean(FertilizerA);
x2<-mean(FertilizerB);
s1<-var(FertilizerA);
s2<-var(FertilizerB);
n1<-length(FertilizerA);
n2<-length(FertilizerB);
t<-(x1-x2)/(sqrt((s1/n1)+(s2/n2)))
alpha <- 0.05;
df <- n1+n2-2;
t_crit <- qt(1 - alpha/2, df);
if(abs(t)<t_crit)
{
sprintf("Do␣not␣reject␣H0␣␣␣μ1 = μ2")
}
if(abs(t)>=t_crit)
{
 sprintf("Reject␣H0␣␣␣μ1 ≠ μ2 ")
}
```

Listing 4.4: R code (Solution example 3)

Example 4. *Sunflower seeds are analyzed for oil content (%).*

Organic:	42.1	41.8	42.0	42.3	41.7	41.9
Conventional:	44.5	45.0	44.2	44.7	44.3	45.1

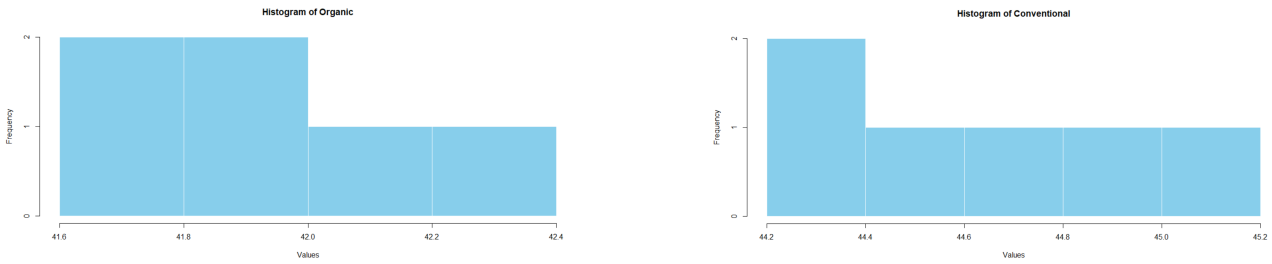
1. *Create* ***histograms*** *and* ***boxplots*** *for both groups.*

2. *Test for* ***equality of variances****.*

```r
# R code (The Histograms and Boxplots for both groups)
Organic<-c(42.1,41.8,42.0,42.3,41.7,41.9);
Conventional<-c(44.5,45.0,44.2,44.7,44.3,45.1);

# Histogram Organic
hist(Organic,
     col = "skyblue",
     border = "white",
     main = "Histogram␣of␣Organic",
     xlab = "Values",
     ylab = "Frequency")

# Boxplot Organic
boxplot(Organic,
        col = "tomato",
```

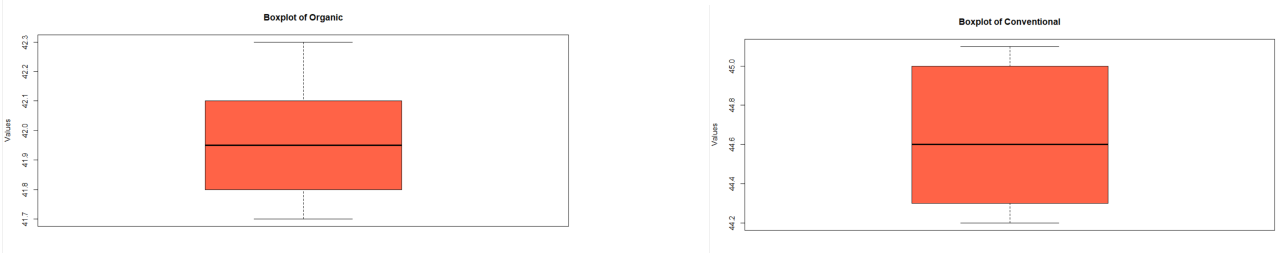
```
        main = "Boxplot␣of␣Organic",
        ylab = "Values")

# Histogram Conventional
hist(Conventional,
     col = "skyblue",
     border = "white",
     main = "Histogram␣of␣Conventional",
     xlab = "Values",
     ylab = "Frequency")

# Boxplot Conventional
boxplot(Conventional,
        col = "tomato",
        main = "Boxplot␣of␣Conventional",
        ylab = "Values")
```

Listing 4.5: R code histograms and boxplots for both groups



(a) histogram Organic (b) histogram Conventional

Figure 4.1: **histograms** for both groups



(a) Boxplot Organic (b) Boxplot Conventional

Figure 4.2: **Boxplots** for both groups

```
# R code (Test the equality of variances)
Conventional<-c(44.5,45.0,44.2,44.7,44.3,45.1);
s1<-var(Organic);
s2<-var(Conventional);

n1<-length(Organic);
n2<-length(Conventional);
if((s1/s2)>0)
{
```

```
F=(s1/s2);
}else{
F=(s2/s1);
}
alpha <- 0.05;
df1 <- n1;
df2 <- n2;
F_crit=qf(1 - alpha, df1, df2);

if(F<F_crit)
{
sprintf("Do␣not␣reject␣H0␣␣␣σ1 = σ2")
}else{
 sprintf("Reject␣H0␣␣␣σ1 ≠ σ2")
}
```

Listing 4.6: R code Test the equality of variances

Chaptre 3

Example 5. *A fish farm tests three water salinity levels (low, medium, high) on juvenile fish weight gain (g/month).*

Low Salinity:	37	44	31
Medium Salinity:	38	39	40
High Salinity:	44	41	49

1. *Analyze the data using one-way ANOVA.*
2. *Interpret whether salinity has a significant effect.*

Chaptre 4

Example 6. *An agronomist studies how* ***plant density (plants/m²)*** *affects* ***potato yield (kg/m²)****. The data collected is as follows:*

Plant Density (X):	2	3	4	5	6
Potato Yield (Y):	3.0	3.8	4.5	4.9	5.1

Tasks:

1. *Compute the* ***Covariance and Correlation coefficient****.*
2. *Estimate the* ***linear regression equation*** $Y = a + bX$*.*

Conclusion

This course has taken students from the fundamentals of statistical description to the practice of statistical inference. They have learned to frame questions, select models that match the design, verify assumptions with diagnostic tools, and report results with the precision expected in professional settings. The tools now in hand include estimation for means, proportions, and variances, hypothesis testing for one and two samples, chi-square procedures for categorical data, one-way and two-way ANOVA for comparing several means, and linear regression for modelling relationships.

Three commitments should guide future work. First, integrity. Plan analyses carefully, document data preparation and code, and present uncertainty transparently rather than relying on single thresholds. Second, relevance. Prioritize effect sizes and confidence intervals that speak to agronomic outcomes, and discuss practical significance alongside statistical significance. Third, reproducibility. Keep a clear record of decisions and assumptions so that colleagues can verify and build upon the analysis.

The techniques introduced here form a foundation for more advanced methods that are highly relevant to agronomy, including generalized linear models for binary and count data, mixed models for multi-location or multi-year trials, and spatial or temporal modelling for precision agriculture. With the discipline and judgment developed in this course, students are prepared to extend their toolkit responsibly. Biostatistics serves the science of sustainable agriculture. The goal is not only correct computation but also sound reasoning that improves decisions in the field and in the laboratory.

Bibliography

[1] Burmeister, L. (1975). Biostatistics in agriculture.

[2] Colquhoun, D. (1971). Lectures on biostatistics: an introduction to statistics with applications in biology and medicine. David Colquhoun.

[3] Gaddis, G. M., Gaddis, M. L. (1990). Introduction to biostatistics: Part 2, descriptive statistics. Annals of Emergency Medicine, 19(3), 309-315.

[4] Gaddis, M. L., Gaddis, G. M. (1990). Introduction to biostatistics: part 6, correlation and regression. Annals of emergency medicine, 19(12), 1462-1468.

[5] Gerstman, B. B., Gerstman, B. B. (2014). Basic biostatistics. Burlington, MA: Jones Bartlett Publishers.

[6] Glaz, B., Yeater, K. M. (2020). Applied statistics in agricultural, biological, and environmental sciences. John Wiley Sons.

[7] Gower, J. C. (1988). Statistics and agriculture. Journal of the Royal Statistical Society Series A: Statistics in Society, 151(1), 179-200.

[8] Härdle, W., Mori, Y., Vieu, P. (Eds.). (2006). Statistical methods for biostatistics and related fields. Springer Science Business Media.

[9] Hatem, G., Zeidan, J., Goossens, M., Moreira, C. (2022). Normality testing methods and the importance of skewness and kurtosis in statistical analysis. BAU Journal-Science and Technology, 3(2), 7.

[10] Hoshmand, R. (2017). Statistical methods for environmental and agricultural sciences. CRC press.

[11] Jobson, J. D. (1991). Multiple linear regression. In Applied multivariate data analysis: Regression and experimental design (pp. 219-398). New York, NY: Springer New York.

[12] Lawal, B. (2014). Applied statistical methods in agriculture, health and life sciences. Springer.

[13] Le, C. T., Eberly, L. E. (2016). Introductory biostatistics. John Wiley Sons.

[14] Lepš, J., Šmilauer, P. (2020). Biostatistics with R: an introductory guide for field biologists. Cambridge University Press.

[15] Looney, S. W. (Ed.). (2002). Biostatistical methods (Vol. 184). University of Louisville School of Medicine, Kentucky: Humana Press.

[16] MacFarland, T. W., Yates, J. M. (2021). Using R for biostatistics. Berlin: Springer.

[17] Nick, T. G. (2007). Descriptive statistics. Topics in biostatistics, 33-52.

[18] Norman, G. R., Streiner, D. L. (2008). Biostatistics: the bare essentials. PMPH USA (BC Decker).

[19] Rao, P. S., Richard, J. (2012). Introduction to biostatistics and research methods. PHI Learning Pvt. Ltd..

[20] Rosner, B. A. (2006). Fundamentals of biostatistics (Vol. 6). Belmont, CA: Thomson-Brooks/Cole.

[21] Sahu, P. K. (2016). Applied statistics for agriculture, veterinary, fishery, dairy and allied fields (pp. 133-194). India:: springer.

[22] Sahu, P. K. (2016). Introduction to Statistics and Biostatistics. In Applied Statistics for Agriculture, Veterinary, Fishery, Dairy and Allied Fields (pp. 1-8). New Delhi: Springer India.

[23] Sokal, R. R., Rohlf, F. J. (1987). Biostatistics. Francise Co, New York, 10, 2088-2092.

[24] Pagano, M., Gauvreau, K., Mattie, H. (2022). Principles of biostatistics. Chapman and Hall/CRC.

[25] Williams, B. (2017). Biostatistics: concepts and applications for biologists. CRC Press.