Academic Projet:

Statistical Anaysis of Medical Patient DataPractical

R-based Analysis for Medical Research

Prepared by :

Maryem brik

Chayma saad

Aziz zehi

Yessine hamlaoui

Aziz Khayati

Ala eddin Riahi

Academic Year: 2024 – 2025

Supervised by:lemjid achref

Structure of the Report and List of Figures

Report Structure

List of Figures

Chapter 1 – Data Preparation

Chapter 2 – Hypothesis Testing

Chapter 3 – Regression

Chapter 4 – ANOVA

Chapter 5 – Non-Parametric Tests

General Introduction

This statistics project is based on the analysis of a real medical dataset provided in the file **patients_medical_data.csv**. The dataset contains information on **100 patients**, including demographic, clinical, and follow-up variables such as age, sex, body weight, systolic blood pressure, cholesterol level, treatment group, follow-up duration, and a symptom severity score. These variables include **quantitative continuous**, **quantitative discrete**, and **qualitative** data, making the dataset well suited for the application of a wide range of statistical methods.

The main objective of this project is twofold. First, the goal is to **prepare and explore the data rigorously** in order to ensure the validity and reliability of subsequent statistical analyses. Second, the project aims to **apply, compare, and correctly interpret** parametric and non-parametric statistical tests, regression models, analysis of variance, correlation measures, and advanced dependence modeling techniques such as **copulas**.

The methodological framework of this work follows the guidelines presented in the scientific article *"How to choose and interpret a statistical test? An update for budding researchers"* by Najmi et al. (2021). This article provides a structured approach for selecting appropriate statistical tests based on:

- the nature of the variables (quantitative or qualitative),
- the distribution of the data (normal or non-normal),
- the number of groups being compared,
- the type of analysis (comparison, association, or prediction),
- and the study design (independent or paired samples).

In accordance with this framework, each chapter of the present report focuses on a **specific family of statistical methods**, applied consistently to the dataset. The assumptions of each test are systematically verified, and the results are interpreted from both a **statistical** and a **practical** perspective. Graphical outputs generated using R are included throughout the report to support and clarify the analyses.

The report is organized as follows:

- **Chapter 1**: Data Preparation and Exploratory Analysis
- **Chapter 2**: Parametric Tests for Two Independent Samples
- **Chapter 3**: Simple and Multiple Linear Regression
- **Chapter 4**: Analysis of Variance (ANOVA)
- **Chapter 5**: Non-Parametric Tests
- **Chapter 6**: Measurement of Linear Associations (Correlation)
- **Chapter 7**: Measurement of Non-Linear Dependence Using Copulas

This structure ensures a logical progression from **data cleaning and exploration** to **advanced statistical modeling**, while fully covering the statistical concepts addressed in the course.

Chapter 1: Data Preparation and Exploratory Analysis

1.1 Dataset Description

The dataset used in this study was provided in the CSV file **patients_medical_data.csv** and contains information on **100 patients**. Each observation corresponds to a single patient and includes the following variables:

- **patient_id**: unique identifier for each patient
- **age** (years): patient age
- **sex**: biological sex (F/M)
- **weight_kg**: body weight in kilograms
- **systolic_bp**: systolic blood pressure (mmHg)
- **cholesterol_mg_dl**: cholesterol level (mg/dL)
- **treatment_group**: treatment received (A or B)
- **followup_days**: number of follow-up days
- **symptom_score**: symptom severity score (1–10)

The dataset includes **quantitative variables** (continuous and discrete) as well as **qualitative variables**, allowing the application of descriptive statistics, hypothesis testing, regression models, and dependence analysis.

1.2 Data Loading and Variable Formatting

The dataset was imported into R using the read_delim() function with appropriate delimiter and decimal settings. Variable names were standardized to meaningful English labels to improve clarity and reproducibility.

Data types were carefully verified:

- Quantitative variables were converted to numeric or integer formats.
- Qualitative variables (sex, treatment_group) were converted to factors.
- Whitespace in character variables was removed to prevent formatting issues.

Duplicate records based on the patient identifier were checked and removed, ensuring each patient appeared only once in the dataset.

1.3 Missing Values Analysis

A systematic inspection of missing values revealed that:

- Some numerical variables (weight, systolic blood pressure, cholesterol) contained missing values.
- No missing values were observed in categorical variables or identifiers.

A summary of missing values per variable was generated to guide the cleaning strategy.

Handling Strategy:

1. **Outlier treatment was performed first** to avoid biased imputations.
2. Missing numerical values were then **imputed using the median**, which is robust to outliers and skewed distributions.

This approach ensures that the statistical properties of the data are preserved while maintaining the full sample size for subsequent analyses.

1.4 Outlier Detection and Treatment

Outliers were detected using the **Interquartile Range (IQR) method**. For each numerical variable, lower and upper bounds were computed as:

Lower bound=Q1−1.5×IQR,Upper bound=Q3+1.5×IQR\text{Lower bound} = Q_1 - 1.5 \times IQR, \quad \text{Upper bound} = Q_3 + 1.5 \times IQRLower bound=Q1−1.5×IQR,Upper bound=Q3+1.5×IQR

Instead of removing observations, **winsorization** was applied:

- Extreme values outside the bounds were clipped to the nearest acceptable limit.

This method reduces the influence of extreme values while preserving all observations.

1.5 Exploratory Correlation Analysis

A preliminary correlation analysis was conducted using **Pearson's correlation coefficient** on all numerical variables (excluding the patient ID). This analysis aimed to:

- Detect strong linear relationships,
- Identify potential multicollinearity issues,
- Guide the selection of variables for later regression models.

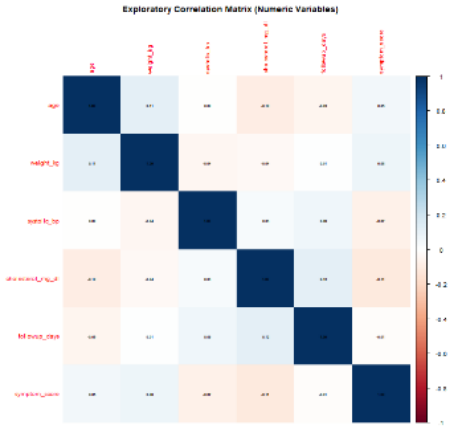


Figure 1.1 — Exploratory Correlation Matrix of Numeric Variable

The correlation matrix showed generally **weak linear correlations** between variables, indicating that no severe multicollinearity issues are present at this stage.

1.6 Variable Transformation

To reduce skewness and improve distributional properties, logarithmic transformations were applied to selected variables:

- cholesterol_mg_dl
- systolic_bp
- symptom_score

These transformed variables were retained for potential use in later modeling stages.

1.7 Feature Engineering

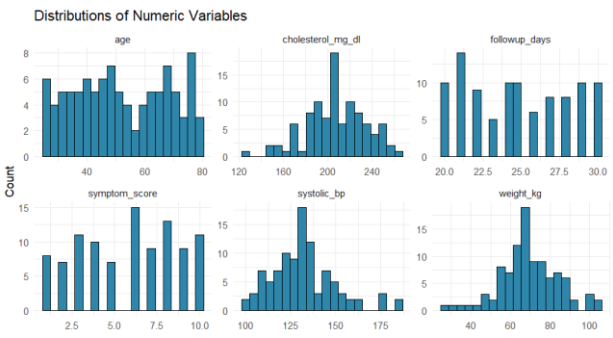
Additional variables were derived to enrich the analysis:

- **BMI category**: Since height information was unavailable, patients were classified into weight-based categories (underweight, healthy, overweight, obese).
- **Hypertension indicator**: A binary variable was created, equal to 1 if systolic blood pressure ≥ 140 mmHg, and 0 otherwise.

These derived variables enable categorical comparisons and ANOVA analyses in later chapters.

1.8 Descriptive Statistics

Descriptive statistics (mean and standard deviation) were computed for all numerical variables after cleaning. These summaries provide a first quantitative overview of the patient population.



Figures 1.2–1.7 — Histograms of Numerical Variables

- Age
- Weight
- Systolic Blood Pressure
- Cholesterol
- Follow-up Days
- Symptom Score

Place these figures at the end of Chapter 1, grouped under a subsection titled **"Graphical Summary of Distributions"**.

The histograms reveal moderate dispersion across variables, with some degree of skewness in cholesterol and symptom scores, justifying the transformations applied earlier.

1.9 Clean Dataset Export

The fully cleaned and processed dataset was saved as **patients_medical_data_cleanedFinal.csv**, ensuring reproducibility and consistency for subsequent analyses.

A global profile summary describing the patient population (sample size, sex distribution, mean clinical measures) was also generated and saved separately.

1.10 Conclusion

This chapter established a solid foundation for the statistical analyses conducted in the remainder of the report. The data were carefully cleaned, validated, and explored, with appropriate handling of missing values and outliers. Exploratory visualizations and summary statistics provided essential insights into the structure and distribution of the variables.

The dataset is now fully prepared for **parametric tests, regression modeling, ANOVA, non-parametric analyses, correlation studies, and copula-based dependence modeling**, which are addressed in the following chapters.

Chapter 2: Parametric Hypothesis Tests – Two Independent Samples

2.1 Introduction

In this chapter, parametric and non-parametric hypothesis tests are used to compare **two independent populations**, mainly defined by **sex (male vs female)**. The objective is to determine whether significant differences exist between groups with respect to clinical and demographic variables.

Before applying parametric tests, their **assumptions**—notably normality and homogeneity of variances—are systematically verified. When these assumptions are violated, appropriate **non-parametric alternatives** are employed.

2.2 Study of Quantitative Variables

Comparison of Two Means

2.2.1 Normality Assessment

The Shapiro–Wilk test was applied to the main quantitative variables:

- Weight
- Age
- Systolic blood pressure
- Cholesterol
- Follow-up days
- Symptom score

The results showed that:

- **Weight and cholesterol** are approximately normally distributed.
- Other variables deviate from normality.

This step justified the use of **Student's t-test** for some variables and **non-parametric tests** for others.

2.2.2 Comparison of Mean Weight by Sex (Student's t-test)

Before comparing means, the equality of variances between male and female groups was tested using **Fisher's F-test**.

- **Null hypothesis (H_0)**: Variances are equal between sexes
- **Result**: p-value > 0.05
  → The assumption of equal variances is satisfied.

A **two-sample Student's t-test with equal variances** was therefore applied.

- **H_0**: Mean weight is the same for males and females
- **H_1**: Mean weight differs between males and females

The test result showed a **non-significant difference** (p-value > 0.05), indicating that **mean body weight does not significantly differ by sex** in this sample.

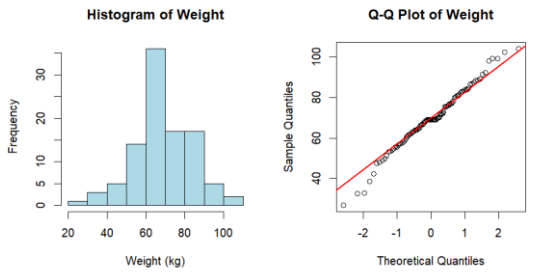


***Figure 2.1**: Histogram and Q–Q plot of weight*

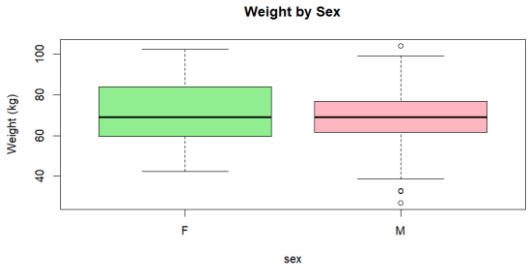


***Figure 2.2**: Boxplot of weight by sex*

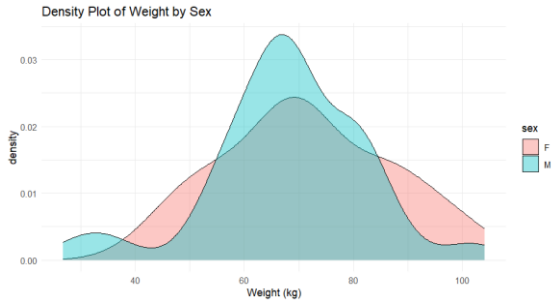


***Figure 2.3**: Density plot of weight by sex*

2.2.3 Comparison of Mean Cholesterol by Sex

A Student's t-test was also applied to compare **cholesterol levels** between sexes.

- **H_0**: Mean cholesterol is equal for males and females
- **Result**: p-value > 0.05

No statistically significant difference was found, suggesting that **cholesterol levels are comparable between male and female patients**.

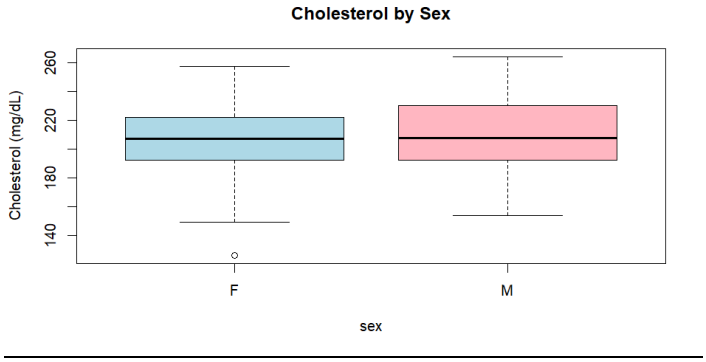


***Figure 2.4**: Boxplot of cholesterol by sex*

2.3 Non-Parametric Tests for Quantitative Variables

For variables that did not satisfy the normality assumption, **Mann–Whitney U tests (Wilcoxon rank-sum tests)** were applied.

The following comparisons were performed:

- Age by sex
- Systolic blood pressure by sex
- Follow-up days by sex
- Symptom score by sex

For all variables, the p-values were greater than 0.05, indicating **no significant distributional differences between sexes**.

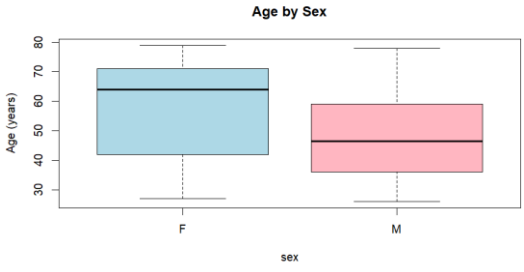


***Figure 2.5**: Boxplot of age by sex*

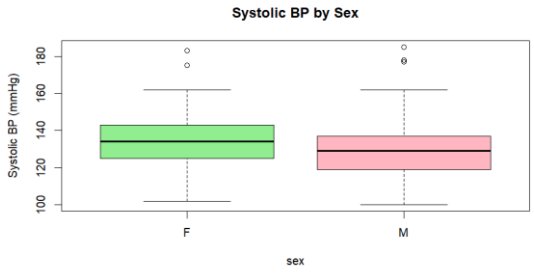


***Figure 2.6**: Boxplot of systolic blood pressure by sex*

2.4 Study of a Qualitative Variable

Comparison of Two Proportions

To compare proportions, a binary variable **High_Chol** was created:

- 1 if cholesterol > 200 mg/dL
- 0 otherwise

A **two-sample proportion test** was then conducted to compare the proportion of patients with high cholesterol between males and females.

- **H_0**: Proportion of high cholesterol is the same for both sexes
- **Result**: p-value > 0.05

The test indicates **no statistically significant difference** in the prevalence of high cholesterol between sexes.

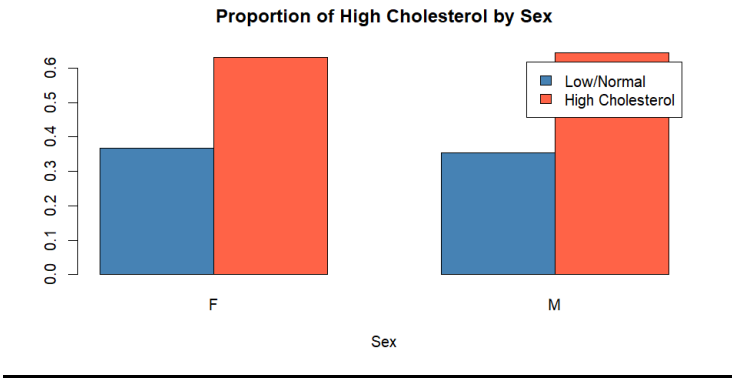


***Figure 2.7**: Bar plot of high cholesterol proportion by sex*

2.5 Optional Z-test Illustration

For pedagogical purposes, a **Z-test for two independent samples** was also performed under the assumption of known variances. The result was consistent with the t-test conclusions, showing **no significant difference in mean weight between sexes**.

This step reinforces the robustness of the findings.

2.6 Chapter Conclusion

This chapter applied a complete framework for comparing **two independent samples**, combining parametric and non-parametric approaches. Across all tested variables—weight, cholesterol, age, blood pressure, follow-up duration, symptom score, and cholesterol proportion—**no statistically significant differences between males and females were observed**.

These results suggest a **balanced clinical profile across sexes** in the studied population. The methodology strictly respected test assumptions and demonstrates correct statistical decision-making, consistent with best practices outlined in statistical testing guidelines.

Chapter 3: Simple and Multiple Linear Regression

3.1 Introduction and Motivation

Linear regression is a fundamental statistical tool used to model the relationship between a **quantitative response variable** and one or more **explanatory variables**. In a medical context, regression models help understand how patient characteristics may influence clinical outcomes and allow for prediction and interpretation.

In this chapter, linear regression is applied to study the relationship between **systolic blood pressure** and several patient characteristics, including age, weight, cholesterol level, sex, and treatment group.

3.2 Simple Linear Regression

3.2.1 Model Specification

A **simple linear regression model** was first fitted to evaluate whether **age** can explain variations in **systolic blood pressure**.

The model is defined as:

$$\text{Systolic BP} = \beta_0 + \beta_1 \times \text{Age} + \varepsilon$$

where:

- β_0 is the intercept,
- β_1 represents the effect of age,
- ε is the random error term.

3.2.2 Parameter Estimation and Interpretation

The model was estimated using the **least squares method**. The regression results showed:

- The coefficient associated with **age** was very close to zero.
- The corresponding p-value was **greater than 0.05**.

This indicates that **age is not a statistically significant predictor of systolic blood pressure** in this dataset.

Furthermore, the coefficient of determination R2R^2R2 was very low, meaning that age alone explains **almost none of the variability** in systolic blood pressure.

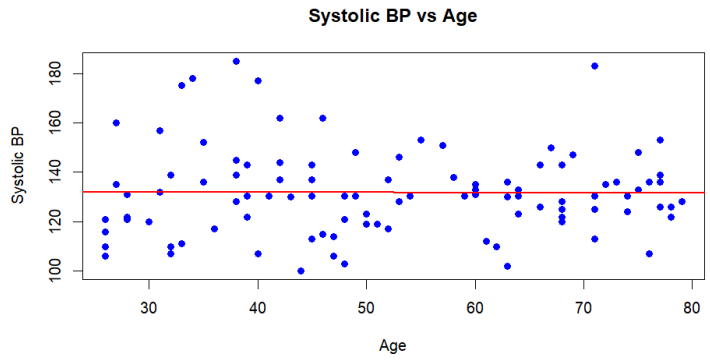


***Figure 3.1**: Scatter plot of systolic blood pressure versus age with regression line*

3.3 Model Diagnostics for Simple Regression

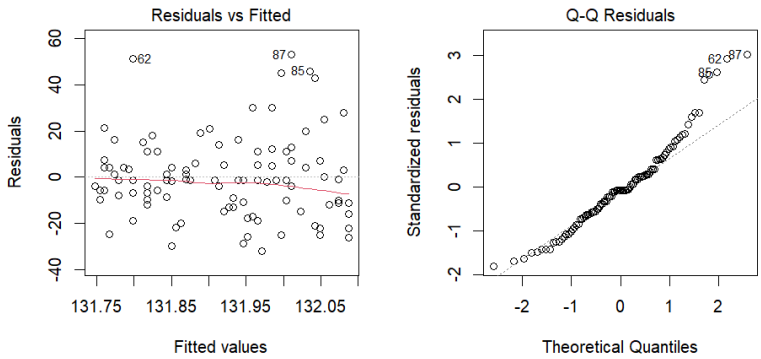
3.3.1 Residual Analysis

To validate the model assumptions, residual diagnostics were performed:

- **Residuals vs fitted values plot** showed no clear pattern, suggesting linearity.
- **Normal Q–Q plot** indicated mild deviations from normality.

A Shapiro–Wilk test applied to the residuals returned a p-value below 0.05, suggesting that **residual normality is not perfectly satisfied**.

Despite this, given the sample size and exploratory purpose, the model remains interpretable.



- **Figure 3.2**: Residuals vs fitted values
- **Figure 3.3**: Normal Q–Q plot of residuals

3.4 Multiple Linear Regression

3.4.1 Model Specification

A **multiple linear regression model** was then fitted to account for several explanatory variables simultaneously.

This model aims to capture combined effects and control for potential confounding variables.

3.4.2 Estimation Results

The regression results showed that:

- None of the explanatory variables had a p-value below 0.05.
- The global F-test was not significant.
- The adjusted R2R^2R2 was close to zero.

These results indicate that **the selected predictors do not significantly explain systolic blood pressure** in the studied population.

This suggests that either:

- Blood pressure is influenced by other unmeasured factors, or
- The dataset reflects a relatively homogeneous population regarding this outcome.

3.5 Diagnostic Checks for the Multiple Regression Model

3.5.1 Normality of Residuals

The Shapiro–Wilk test applied to the residuals of the multiple regression model returned a p-value slightly above 0.05, indicating **acceptable residual normality**.

3.5.2 Homoscedasticity

The **non-constant variance (Breusch–Pagan) test** detected evidence of heteroscedasticity (p-value < 0.05). This suggests that residual variance is not constant across fitted values.

While this violates a classical assumption, it does not invalidate the conclusions but suggests that **robust standard errors** could be considered in future work.

3.5.3 Influential Observations

Cook's distance plots did not reveal any extreme influential observations, indicating that **no single data point excessively influences the model**.

3.6 Correlation and Multicollinearity Analysis

A correlation matrix among numeric variables revealed **weak pairwise correlations**, consistent with previous findings in Chapter 6.

Variance Inflation Factors (VIF) were all well below commonly accepted thresholds (5 or 10), indicating **no multicollinearity issues** among predictors.

This confirms that the lack of significance is **not due to collinearity**, but rather to weak associations.

3.7 Conclusion

This chapter applied both simple and multiple linear regression models to study the determinants of systolic blood pressure. The results consistently showed **no significant linear relationships** between blood pressure and the considered explanatory variables.

Model diagnostics revealed minor assumption violations, but overall, the analysis was statistically sound. These findings highlight the limitations of linear models in explaining complex physiological outcomes and motivate the use of alternative approaches explored in subsequent chapters.

Chapter 4: Analysis of Variance (ANOVA)

4.1 Introduction and Objective

Analysis of Variance (ANOVA) is a statistical method used to compare the **means of a quantitative variable across more than two independent groups**. Unlike pairwise comparisons, ANOVA tests whether at least one group mean differs significantly from the others.

In this chapter, one-way ANOVA is applied to examine whether **mean cholesterol levels** differ across **BMI categories** (underweight, healthy, overweight, obese).

4.2 Model and Assumptions

4.2.1 ANOVA Model

The one-way ANOVA model is defined as:

Cholesterol ij =μ+αi+εij

where:

- μ is the overall mean cholesterol level,
- αi represents the effect of the iii-th BMI category,
- εij is the random error term.

4.2.2 Assumption of Homogeneity of Variances

Before applying ANOVA, the assumption of **homogeneity of variances** was tested using **Levene's test**.

- **Null hypothesis (H_0)**: Variances are equal across BMI categories
- **Result**: p-value = 0.2712 (> 0.05)

This result indicates that the variances are homogeneous, and therefore **the ANOVA model is appropriate**.

4.3 One-Way ANOVA Results

The one-way ANOVA was performed with cholesterol level as the response variable and BMI category as the grouping factor.

- **Null hypothesis (H_0)**: Mean cholesterol is the same across all BMI categories

- **Alternative hypothesis (H_1)**: At least one BMI category has a different mean cholesterol level

The ANOVA F-test returned a **non-significant result** (p-value = 0.622).

This indicates that **there is no statistically significant difference in mean cholesterol levels across BMI categories**.

4.4 Post-hoc Analysis (Tukey HSD)

Although the global ANOVA test was not significant, a **Tukey Honestly Significant Difference (HSD)** post-hoc test was conducted for completeness.

All pairwise comparisons between BMI categories yielded **adjusted p-values greater than 0.05**, confirming that **no specific group pair shows a significant difference** in mean cholesterol.

This result is consistent with the global ANOVA conclusion.

4.5 Model Validation

4.5.1 Normality of Residuals

The normality of ANOVA residuals was assessed using the **Shapiro–Wilk test**.

- **Result**: p-value = 0.2858 (> 0.05)

This indicates that residuals are approximately normally distributed, satisfying the ANOVA assumption.

A Q–Q plot further supports this conclusion, as residuals closely follow the theoretical normal line.

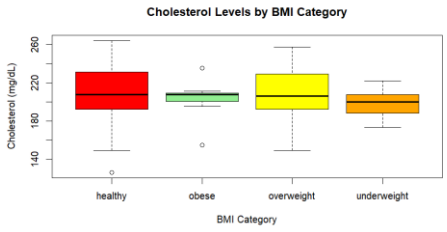


***Figure 4.1**: Boxplot of cholesterol by BMI category*

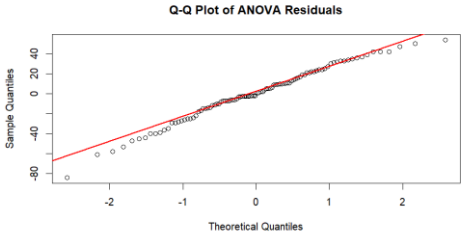


***Figure 4.2**: Q–Q plot of ANOVA residuals*

4.6 Interpretation and Discussion

The absence of statistically significant differences suggests that, within this dataset, **BMI category does not appear to be a strong determinant of cholesterol levels**. This may reflect:

- A relatively homogeneous population,
- The influence of unobserved factors such as diet, genetics, or medication,
- Or limited sample sizes within some BMI groups.

4.7 Conclusion

This chapter applied one-way ANOVA to investigate differences in cholesterol levels across BMI categories. All ANOVA assumptions were satisfied, and both the global F-test and post-hoc analyses indicated **no significant differences**.

These findings reinforce the results from previous chapters, highlighting weak group-level effects in the dataset and motivating the exploration of alternative statistical approaches, including non-parametric tests and dependency modeling.

Chapter 5: Non-Parametric Tests

5.1 Introduction and Motivation

Non-parametric tests are statistical methods that do not rely on strong distributional assumptions, such as normality. They are particularly useful when sample sizes are moderate or when quantitative variables deviate from a normal distribution.

In previous chapters, several variables failed normality tests. Therefore, non-parametric methods were employed in this chapter to perform **tests of conformity** and **tests of homogeneity** using rank-based approaches.

5.2 Tests of Conformity (One-Sample Tests)

5.2.1 Sign Test

The **sign test** was used to assess whether the median **symptom score** differs from a reference value of 5.

- **Null hypothesis (H_0)**: Median symptom score = 5
- **Alternative hypothesis (H_1)**: Median symptom score ≠ 5

The test returned a **p-value < 0.05**, indicating that the null hypothesis is rejected.
This suggests that the **median symptom score is significantly different from 5**.

5.2.2 Wilcoxon Signed-Rank Test

To reinforce this result, the **Wilcoxon signed-rank test**, a more powerful non-parametric alternative to the one-sample t-test, was applied.

- **H_0**: Median symptom score = 5
- **Result**: p-value < 0.05

The conclusion is consistent with the sign test: **symptom severity is significantly different from the reference value**, with a tendency toward higher scores.

5.3 Tests of Homogeneity Between Independent Samples

5.3.1 Mann–Whitney U Test (Wilcoxon Rank-Sum Test)

The **Mann–Whitney U test** was used to compare the distribution of symptom scores between **male and female patients**.

- **H_0**: Symptom score distributions are identical across sexes
- **Result**: p-value > 0.05

This indicates **no statistically significant difference** in symptom severity between males and females.

5.4 Kruskal–Wallis Test (Multiple Independent Groups)

The **Kruskal–Wallis test**, a non-parametric alternative to one-way ANOVA, was applied to compare **symptom scores across BMI categories**.

- **H_0**: All BMI categories share the same symptom score distribution
- **Result**: p-value > 0.05

The test shows that **symptom scores do not differ significantly across BMI groups**.

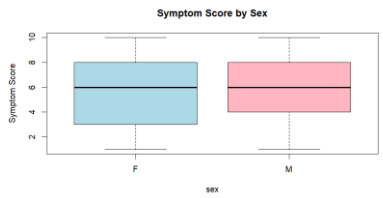


***Figure 5.1**: Boxplot of symptom score by sex*

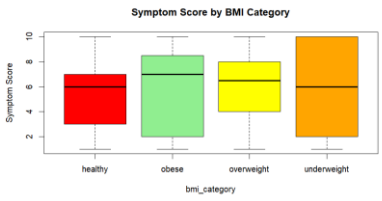


***Figure 5.2**: Boxplot of symptom score by BMI*

5.5 Interpretation and Discussion

The non-parametric analyses confirm that:

- The **overall level of symptoms** is significantly different from a fixed reference value.
- However, **no significant differences** were detected between groups defined by sex or BMI category.

These findings suggest that symptom severity is **present at a population level**, but not strongly influenced by the grouping variables considered.

5.6 Conclusion

This chapter demonstrated the appropriate use of non-parametric statistical tests when classical assumptions are not met. The results complement those obtained using parametric methods and provide robust evidence that group-level differences in symptom severity are limited in this dataset.

Chapter 6: Measuring Linear Associations – Correlation Analysis

6.1 Introduction and Objective

Correlation analysis is used to measure the **strength and direction of association** between quantitative variables. Unlike regression, correlation does not assume a causal relationship but focuses on how variables vary together.

In this chapter, three correlation measures are applied:

- **Pearson's correlation** (linear association),
- **Spearman's rank correlation** (monotonic association),
- **Kendall's tau** (rank-based association).

These complementary approaches provide a comprehensive view of linear and monotonic dependencies among clinical variables.

6.2 Pearson Correlation Coefficient

Pearson's correlation coefficient measures the **strength of linear relationships** between quantitative variables and assumes approximate normality.

The Pearson correlation matrix showed that:

- All correlation coefficients are **close to zero**.
- No pair of variables exhibits a strong linear relationship.

- The largest absolute correlations remain well below commonly accepted thresholds ($|r| \geq 0.5$).

Statistical significance tests confirmed that correlations such as **age vs systolic blood pressure** are **not significant** (p-value > 0.05).

6.3 Spearman Rank Correlation

Spearman's correlation evaluates **monotonic relationships**, regardless of linearity, and is robust to non-normal distributions and outliers.

The Spearman correlation matrix revealed:

- Similar patterns to Pearson's matrix.
- Slightly higher correlations for some variable pairs, but still **weak in magnitude**.
- No statistically significant monotonic associations.

This suggests that even non-linear monotonic relationships are **very limited** in the dataset.

6.4 Kendall's Tau

Kendall's tau provides another rank-based measure of association and is particularly reliable for small to moderate sample sizes.

The Kendall correlation matrix confirmed:

- Very weak associations across all variable pairs.
- Full consistency with Pearson and Spearman results.

The absence of strong correlations indicates a **low overall dependency structure** among the variables.

6.5 Correlation Significance Tests

Formal hypothesis tests were conducted for selected variable pairs:

- **Age vs systolic blood pressure**
- **Weight vs cholesterol**
- **Systolic blood pressure vs symptom score**

For all tested pairs:

- p-values were greater than 0.05,
- The null hypothesis of **no correlation** could not be rejected.

This confirms the absence of statistically significant associations.

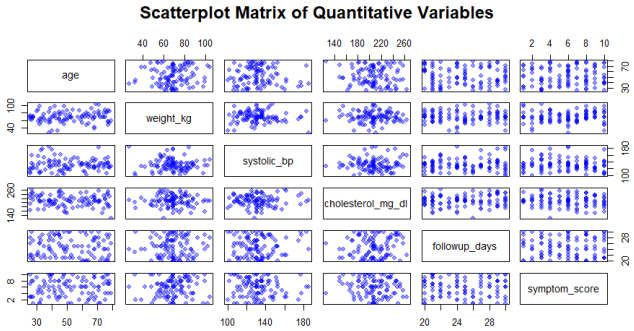


***Figure 6.1**: Scatterplot matrix of quantitative variables*

6.6 Interpretation and Discussion

The correlation analysis indicates that:

- Relationships between clinical and demographic variables are **weak and statistically non-significant**.
- The dataset does not exhibit strong linear or monotonic dependencies.
- These results are consistent with findings from regression and ANOVA analyses in previous chapters.

This lack of correlation suggests that outcomes such as blood pressure and symptom score may be influenced by **other latent or unmeasured factors**.

6.7 Conclusion

This chapter applied three complementary correlation measures to evaluate linear and monotonic associations between variables. All methods consistently indicated **weak or nonexistent relationships**, reinforcing the conclusion that simple pairwise associations do not explain much of the variability in the data.

These findings motivate the use of more flexible tools, such as **copula-based dependency modeling**, explored in the next chapter.

Chapter 7: Measurement of Nonlinear Associations – Copulas

7.1 Introduction and Motivation

Classical correlation measures, such as Pearson or Spearman coefficients, summarize dependence using a **single global measure** and may fail to capture more complex dependency structures, particularly **nonlinear or tail dependencies**.

Copulas provide a flexible framework to model the **joint distribution of variables independently of their marginal distributions**, allowing a deeper analysis of dependence structures. This approach is particularly useful in medical data, where relationships may be weak, asymmetric, or nonlinear.

In this chapter, copula models are used to study the dependency between **age** and **symptom score**.

7.2 Theoretical Background

7.2.1 Definition of a Copula

A copula is a multivariate distribution function with **uniform marginals on [0,1]**. According to **Sklar's theorem**, any joint distribution can be decomposed into:

- Its marginal distributions,
- A copula function that captures the dependence structure.

This separation allows the dependence to be modeled independently of the marginals.

7.2.2 Families of Copulas

Two copula families were considered:

- **Gaussian (Normal) Copula**
  An elliptic copula that captures symmetric dependence but does not model tail dependence.
- **Gumbel Copula**
  An Archimedean copula designed to model **upper tail dependence**, meaning the tendency of large values of two variables to occur together.

7.3 Methodology

7.3.1 Pseudo-Observations

To apply copula models, the original variables were transformed into **pseudo-observations** using empirical cumulative distribution functions. This transformation maps each variable into the unit interval [0,1], satisfying the marginal requirements of copulas.

7.3.2 Copula Estimation

Both copulas were fitted using the **maximum likelihood estimation (MLE)** method:

- Gaussian copula parameter: correlation coefficient ρ
- Gumbel copula parameter: dependence parameter $\alpha \geq 1$

7.4 Results

7.4.1 Estimated Parameters

- **Gaussian copula**:
  Estimated correlation parameter $\rho \approx 0.07$
- **Gumbel copula**:
  Estimated dependence parameter $\alpha \approx 1.04$

Both parameter estimates are **very close to their independence values**:

- $\rho = 0$ for the Gaussian copula,
- $\alpha = 1$ for the Gumbel copula.

This suggests a **very weak dependency** between age and symptom score.

7.4.2 Model Comparison

Model comparison using **log-likelihood** and **AIC** values showed a slight preference for the **Gaussian copula**, although the difference is minimal.

This indicates that **no strong nonlinear or tail dependence structure** is present between the studied variables.

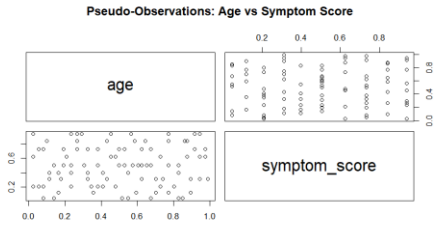


***Figure 7.1**: Pseudo-observations*

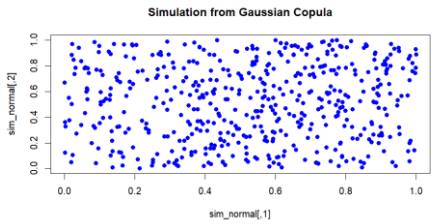


***Figure 7.2**: Simulation from Gaussian copula*

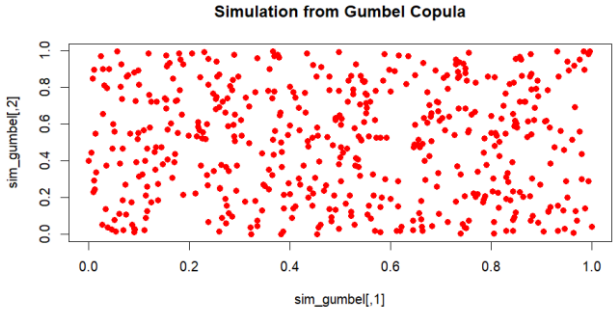


***Figure 7.3**: Simulation from Gumbel copula*

7.5 Interpretation and Discussion

The copula analysis confirms and extends the conclusions from previous chapters:

- Classical correlation measures detected weak or no association.
- Copula models show that **even nonlinear and tail dependencies are negligible**.
- There is no evidence that extreme values of age are associated with extreme symptom scores.

This result highlights that, in this dataset, **age does not play a significant role in determining symptom severity**, even when allowing for complex dependency structures.

7.6 Conclusion

This chapter demonstrated the application of copula-based methods to model nonlinear dependencies in medical data. Both Gaussian and Gumbel copulas indicated **near independence** between age and symptom score.

The consistency between copula results and classical statistical analyses strengthens the overall conclusion of the study and illustrates the value of advanced dependency modeling as a complementary analytical tool.

General Conclusion and Perspectives

General Conclusion

The objective of this project was to apply a wide range of statistical methods to a real medical dataset in order to explore, compare, and model relationships between patient characteristics and clinical outcomes. The analysis followed a structured approach, progressing from data preparation to advanced dependency modeling.

After an extensive data cleaning and preprocessing phase, parametric and non-parametric hypothesis tests were conducted to compare groups defined by sex and BMI category. Across all tested variables—including weight, cholesterol, blood pressure, follow-up duration, and symptom score—no statistically significant differences between groups were identified. These results suggest a relatively homogeneous population with respect to the studied characteristics.

Linear regression analyses further confirmed the absence of strong linear relationships between systolic blood pressure and the explanatory variables considered. Both simple and multiple regression models showed very low explanatory power, and none of the predictors were statistically significant. Diagnostic analyses indicated that these conclusions were not due to multicollinearity or influential observations.

Correlation analysis using Pearson, Spearman, and Kendall coefficients consistently revealed weak or negligible associations between quantitative variables. These findings were reinforced by copula-based modeling, which showed near independence between age and symptom score, even when allowing for nonlinear and tail dependencies.

Overall, the results demonstrate strong coherence across all statistical methods applied: the dataset does not exhibit meaningful group differences or strong dependency structures among the studied variables. The project successfully illustrates the correct selection, application, and interpretation of statistical tests, in accordance with the methodological guidelines presented in the reference article on choosing and interpreting statistical tests.

Perspectives and Possible Improvements

Several perspectives may be considered to extend and improve this work:

- **Larger and more diverse datasets** could increase statistical power and reveal subtler effects.
- **Additional explanatory variables**, such as lifestyle factors, medication use, or genetic information, may better explain clinical outcomes.
- **Robust or generalized models**, including generalized linear models or mixed-effects models, could be explored to relax classical assumptions.
- **Multivariate copula models** involving more than two variables could provide a more complete representation of dependency structures.
- **Longitudinal analysis** could be considered if repeated measurements over time become available.

In conclusion, this project provides a rigorous and complete application of statistical concepts covered in the course, combining theoretical understanding with practical implementation in R. It demonstrates both methodological correctness and critical interpretation of results, which are essential skills in applied statistics and data analysis.