People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Constantine 1 University-Frères Mentouri
Faculty of Exact Sciences
Department of Mathematics



Statistics applied to the health domain

Course Notes

Master 1 – Applied Statistics

Prepared by:
Dalel ZERDAZI

Academic Year: 2025–2026

Contents

1 Introduction to Epidemiology and Health Indicators **3**
1.1 General Introduction . . . 3
1.2 Objectives of Epidemiology . . . 3
1.3 Applied Reflection: Mini-Quiz on Health Statistics . . . 4
1.4 Types of Health Indicators . . . 6
1.4.1 Demographic Indicators . . . 6
1.4.2 Morbidity Indicators . . . 6
1.4.3 Mortality Indicators . . . 6
1.4.4 Composite Indicators . . . 6
1.5 Prevalence and Incidence . . . 7
Exercise: Prevalence and Incidence . . . 7
Practical Work 1: Prevalence and Incidence in R . . . 8
1.6 Measuring Mortality . . . 11
Exercise: Comparing Mortality . . . 12
Practical Work 2: Comparing Mortality in R . . . 13
1.7 Chapter Conclusion . . . 15

2 Organization and Conduct of a Health Survey **16**
2.1 Introduction . . . 16
2.2 Conceptual Foundations of Health Surveys . . . 16
2.2.1 Defining the Objective and Research Question . . . 16
2.2.2 Key Definitions and Notations . . . 17
2.2.3 Defining the Study Population . . . 18
2.2.4 Conceptual Model of Variables . . . 18
2.2.5 From Concept to Measurement . . . 19
2.2.6 Ethical and Practical Considerations . . . 20
2.2.7 Estimation of Population Prevalence under Simple Random Sampling 20
Exercise: Estimation of Hypertension Prevalence . . . 21
Practical Work 3: Estimation of Hypertension Prevalence in R . . . 22
2.3 Choice of Study Design . . . 23
2.3.1 Cross-Sectional Surveys . . . 24
2.3.2 Cohort Studies . . . 26
2.3.3 Case–Control Studies . . . 28
2.3.4 Randomized Controlled Trials (RCTs) . . . 31
Exercise: Comparing Study Designs . . . 33
Practical Work 4: Comparing Study Designs in R . . . 34
2.4 Chapter Conclusion . . . 39

3 Standardization Methods in Epidemiology **40**
3.1 Introduction . . . 40
3.2 Why Standardize? . . . 40
3.3 Key Definitions . . . 40
3.4 Direct Standardization . . . 41
3.4.1 Concept . . . 41
3.4.2 Formula . . . 42
3.4.3 Advantages and Limitations . . . 43
3.5 Indirect Standardization . . . 44
3.5.1 Concept . . . 44
3.5.2 Formula . . . 44
3.5.3 Advantages and Limitations . . . 47
3.6 Comparison of Methods . . . 47
Exercise: Direct and Indirect Standardization of Mortality Rates . . . 48
Practical Work 5: Direct and Indirect Standardization of Mortality Rates in R . 50
3.7 Chapter Conclusion . . . 53

4 Introduction to Logistic Regression **54**
4.1 Motivation and Context . . . 54
4.2 The Logistic Regression Model: Theoretical Foundations . . . 55
4.2.1 Binary Outcomes and the Bernoulli Model . . . 55
4.2.2 The Logistic Function and the Logit Link . . . 56
4.2.3 Why the Logit? . . . 56
4.2.4 The Logistic Regression Equation . . . 56
4.2.5 Likelihood and Maximum Likelihood Estimation . . . 57
4.2.6 Interpretation Through the Odds Ratio . . . 59
4.2.7 Summary of the Theoretical Framework . . . 60
Exercise: Logistic Regression on a Tiny Dataset . . . 60
Practical Work 6: Logistic Regression on a Tiny Dataset in R . . . 63
Exercise: Logistic Regression on Smoking Data . . . 65
Practical Work 7: Logistic Regression on Smoking Data in R . . . 67
Exercise: Maximum Likelihood Estimation in logistic regression . . . 69
Practical Work 8: Maximum Likelihood Estimation in logistic regression in R . 72
4.3 Chapter Conclusion . . . 74

5 Practical Case Studies: Epidemiological and Medico-Economic Analyses **75**
5.1 General Framework for Applied Case Studies . . . 75
5.2 Case Study 01: Multivariable Logistic Regression in Epidemiology . . . 75
5.3 Case Study 02: Effect Modification and Stratified Interpretation . . . 79
5.4 Case Study 03: Count Data and Overdispersion . . . 81
5.5 Case Study 04: Medico-Economic Cost-Effectiveness Analysis . . . 83
5.6 Chapter Conclusion . . . 85

General Introduction

Applied statistics in the health sciences constitutes a fundamental tool for understanding, analyzing, and interpreting health phenomena within populations. It enables the transformation of observations collected from surveys, clinical studies, or health databases into scientifically meaningful information that can support research, medical decision-making, and public health policy.

This course, delivered within the first-semester methodological training unit, aims to provide students with the conceptual, methodological, and computational foundations required for statistical analysis in health sciences. Its primary objective is to develop an integrated understanding of statistical reasoning applied to epidemiological and medico-economic problems, from study design to interpretation of results.

By the end of this course, students will be able to design questionnaires and structure survey protocols, select appropriate study designs, identify relevant health indicators, and analyze collected data using statistical software, particularly `R`. They will thus be capable of applying quantitative methods adapted to public health studies, including risk analysis, population comparisons, modeling relationships between variables, and interpreting findings in real-world health contexts.

The course is organized progressively. It begins with an introduction to the fundamental concepts of epidemiology and health indicators, followed by a presentation of survey organization and study design principles. Methods for measuring risk and techniques of standardization are then introduced to ensure valid comparisons between populations. The course continues with logistic regression, a central tool for multivariable analysis of health data. Finally, several applied case studies illustrate the practical use of statistical methods in epidemiological and medico-economic contexts.

Practical sessions complement the theoretical instruction by allowing students to implement statistical methods using the `R` software environment and real datasets obtained from public databases or locally collected surveys. This hands-on approach promotes active learning, reproducibility of analyses, and the development of operational skills directly applicable to professional practice or research.

This course assumes prior knowledge of basic statistics, probability, descriptive measures, and hypothesis testing, as well as introductory skills in data handling using `R`. Familiarity with fundamental epidemiological concepts such as prevalence, incidence, and risk measures is also recommended. The aim is to consolidate these prerequisites while integrating them into a coherent applied analytical framework.

Overall, this document is designed as a structured pedagogical pathway that connects theoretical foundations with practical applications, enabling students to use statistical tools rigorously, critically, and effectively in the field of health sciences.

Chapter 1

Introduction to Epidemiology and Health Indicators

1.1 General Introduction

Epidemiology is the scientific discipline that studies the distribution, frequency, and determinants of health problems in human populations. It forms one of the pillars of modern public health and helps answer essential questions such as:

- How many people are affected by a given disease?
- What are the causes or associated risk factors?
- Which prevention and control strategies are effective?

Health indicators are measurement tools that translate the health status of a population into numerical data. They allow:

- Describing the health situation,
- Identifying public health priorities,
- Evaluating the impact of health interventions and policies.

Definition (WHO): *"Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems."*

1.2 Objectives of Epidemiology

The main objectives of epidemiology are:

1. **Describe**: measure the frequency of diseases and their distribution by age, sex, location, or time.
2. **Analyze**: identify causes and risk factors.
3. **Evaluate**: assess the effectiveness of interventions (vaccination, prevention campaigns, health policies).

4. **Prevent and act**: provide evidence-based recommendations to improve population health.

Example: During the COVID-19 pandemic, epidemiology helped to:

- Describe the spread of the epidemic (number of cases, mortality rates),
- Identify risk factors (age, comorbidities),
- Assess the effectiveness of vaccines and preventive measures.

1.3 Applied Reflection: Mini-Quiz on Health Statistics

To strengthen understanding of the importance and applications of statistics in health sciences, the following mini-quiz encourages critical thinking about data quality, study design, and interpretation in public health.

Mini-Quiz: Thinking Statistically about Health

The following questions aim to encourage students to think critically about data, study design, and statistical interpretation in public health contexts.

1. A city reports more flu cases this year. What information would you need to understand this increase? *Hints:* Population size, weather conditions, vaccination coverage, mobility of residents.

2. An article claims: "Drinking coffee reduces the risk of heart disease." How could you check if this statement is credible? *Hints:* Study source, number of participants, control for smoking or diet.

3. You organize a survey on students' eating habits. How would you select participants to obtain reliable results? *Hints:* Representative sample (different majors, ages), avoid only close friends, ensure anonymity.

4. Why is it risky to draw a conclusion from one person's experience? *Hints:* Not representative, coincidence, individual variability.

5. Ice cream sales and drowning cases both rise together. What can you conclude? *Hints:* Correlation does not imply causation. Likely common factor: hot weather.

6. A hospital wants to know if a new drug works better than the old one. What should be compared? *Hints:* Two comparable groups, same follow-up, measure before/after.

7. Why is anonymity important when collecting health data? *Hints:* Protect privacy, encourage honesty, ensure ethical compliance.

8. A survey includes only social media users. What problem might arise? *Hints:* Selection bias — exclusion of people without internet access.

9. "Pollution increases asthma attacks." What other factors should you consider before concluding? *Hints:* Seasonal allergies, smoking, weather.

10. Why measure a patient's temperature several times? *Hints:* Reduce random errors, detect variations.

11. Some people do not respond to a survey. What impact can this have? *Hints:* Non-response bias — less representative results.

12. You compare average weight between two student groups. Why must the scales be properly calibrated? *Hints:* Avoid systematic errors affecting all measurements.

13. A new screening test detects 95% of patients. What questions remain? *Hints:* False positives, cost, accessibility, comparison with existing tests.

14. If a chart shows only the mean, why is it useful to know the range (min and max)? *Hints:* The mean alone can hide large variability.

15. In a sleep study, why record diet or physical activity? *Hints:* These variables may influence sleep and confound results.

16. A health center wants to know if a new prevention poster reduces flu cases. What should be compared? *Hints:* Number of cases before/after, same season, same weather.

17. "People who walk 10,000 steps a day live longer." What alternative explanations might exist? *Hints:* Walkers may eat better, have less stress, or be fitter.

18. In a sleep survey, some people overestimate their rest time. What problem does this cause? *Hints:* Measurement bias — less reliable data, distorted conclusions.

19. A study says "living near a park lowers blood pressure." How can we ensure this is not simply linked to income? *Hints:* Adjust for socioeconomic level, compare similar-income neighborhoods.

20. Why repeat the same survey several years in a row? *Hints:* Observe trends, detect changes, measure policy effects.

These questions highlight how statistical reasoning supports critical thinking and sound interpretation of health data.

1.4 Types of Health Indicators

Health indicators are essential tools that summarize the health status of a population and enable comparisons across regions, time periods, and demographic groups. They guide public health decisions, help identify priorities, and evaluate the impact of interventions.

1.4.1 Demographic Indicators

Demographic indicators describe the size, structure, and dynamics of a population.

- **Birth rate:** number of live births per 1,000 inhabitants per year. It reflects fertility and demographic renewal.
- **Life expectancy:** average number of years a newborn is expected to live if current mortality rates remain constant.

1.4.2 Morbidity Indicators

Morbidity indicators measure the frequency and distribution of diseases and other health conditions.

- **Prevalence:** proportion of individuals affected by a disease at a given time.
- **Incidence:** number of new cases occurring in a population during a given period.
- **Attack rate:** proportion of individuals who contract a disease during a specific epidemic or outbreak.

1.4.3 Mortality Indicators

Mortality indicators reflect the frequency of death in a population and are key to assessing health outcomes.

- **Crude death rate (CDR):** number of deaths per 1,000 inhabitants per year.
- **Infant mortality rate:** number of deaths before one year of age per 1,000 live births. It is a sensitive measure of healthcare quality and socioeconomic conditions.
- **Cause-specific mortality rate:** number of deaths due to a specific cause (e.g., lung cancer) per 100,000 inhabitants.
- **Maternal mortality rate:** number of deaths of women from pregnancy-related causes per 100,000 live births.

1.4.4 Composite Indicators

Composite indicators combine several dimensions of health and well-being into a single measure.

- **DALYs** (*Disability-Adjusted Life Years*): years of life lost due to premature death or disability.
- **QALYs** (*Quality-Adjusted Life Years*): years of life adjusted by the quality or utility of those years.

Limitation of Single Indicators:

Prevalence and incidence are essential measures of disease frequency, but they alone cannot capture the overall severity or social impact of health problems. For example, a rare but fatal cancer may show low prevalence yet cause a major health burden. This limitation highlights the need to interpret these indicators carefully and, when necessary, complement them with other measures such as mortality or composite indicators (e.g., DALYs, QALYs).

This naturally leads to a more detailed exploration of these two fundamental measures — **prevalence** and **incidence** — in the following section.

1.5 Prevalence and Incidence

Prevalence and incidence are two fundamental concepts in epidemiology used to describe disease frequency in a population.

- **Prevalence:**
$$P = \frac{\text{Number of existing cases at a given time}}{\text{Population at risk}}$$
It measures the proportion of individuals currently affected by the disease.
- **Incidence:**
$$I = \frac{\text{Number of new cases during a period}}{\text{Population at risk at the beginning of the period}}$$
It measures the rate at which new cases occur within a specified time frame.

Relationship between the two:

$$\text{Prevalence} \approx \text{Incidence} \times \text{Average duration of the disease}$$

This relationship implies that a disease can have:

- High prevalence if it is chronic and lasts long (e.g., diabetes),
- Or low prevalence if it is acute or rapidly fatal (e.g., influenza).

Example: A chronic disease such as diabetes shows high prevalence because patients often live many years with it, even if the annual incidence is relatively low.

Exercise: Prevalence and Incidence

Problem Statement

In a city of 10,000 inhabitants:

- 200 people are diabetic on January 1st.
- 50 new cases occur during the year.

1. Calculate the prevalence on January 1st.
2. Calculate the annual incidence.
3. What is the prevalence on December 31st?
4. Interpret.

Solution

1. **Prevalence on January 1st:**

   Prevalence measures the proportion of individuals who have the disease at a given point in time (point prevalence).

$$P_{\text{Jan 1}} = \frac{\text{Number of existing cases}}{\text{Total population}} \times 100 = \frac{200}{10\,000} \times 100 = 2\%$$

   Interpretation: On January 1st, 2% of the population is already diabetic.

2. **Annual incidence:**

   Incidence measures the proportion of new cases among individuals at risk during a specific period.

$$I_{\text{annual}} = \frac{\text{Number of new cases}}{\text{Population at risk}} \times 100 = \frac{50}{10\,000 - 200} \times 100 = \frac{50}{9\,800} \times 100 \approx 0.51\%$$

   Interpretation: During the year, approximately 0.51% of the population who were initially free of diabetes developed the disease.

3. **Prevalence on December 31st:**

   End-of-year prevalence includes both existing and new cases:

$$P_{\text{Dec 31}} = \frac{\text{Existing cases} + \text{New cases}}{\text{Total population}} \times 100 = \frac{200 + 50}{10\,000} \times 100 = 2.5\%$$

   Interpretation: By the end of the year, 2.5% of the population is diabetic.

4. **Overall interpretation:**

   - The chronic disease shows a slow but steady increase in prevalence.
   - While the incidence is relatively low, it contributes to a growing number of cases over time.
   - Monitoring both prevalence and incidence is essential to understand the disease burden and to plan public health interventions.

Practical Work 1: Prevalence and Incidence in R

Step 1 — Basic Data Definition

```r
population <- 10000
existing_cases <- 200
new_cases <- 50
population_at_risk <- population - existing_cases
```

Explanation:

- **population:** total number of individuals in the population (10,000 persons).

- **existing_cases:** number of existing cases at the beginning of the year.
- **new_cases:** number of new cases occurring during the year.
- **population_at_risk**: individuals who are still at risk of developing the disease, i.e. the total population minus the existing cases.

Thus,

$$\text{Population at risk} = 10\,000 - 200 = 9\,800.$$

Step 2 — Epidemiological Calculations

```
prevalence_init <- existing_cases / population
incidence <- new_cases / population_at_risk
prevalence_final <- (existing_cases + new_cases) / population
```

Explanation:

- **Initial Prevalence:**

$$\text{Prevalence}_{\text{init}} = \frac{200}{10\,000} = 0.02 = 2\%.$$

This represents the proportion of individuals already ill at the beginning of the year.

- **Annual Incidence:**

$$\text{Incidence} = \frac{50}{9\,800} = 0.0051 \approx 0.51\%.$$

This measures the frequency of new cases among the population at risk.

- **Final Prevalence:**

$$\text{Prevalence}_{\text{final}} = \frac{200 + 50}{10\,000} = 0.025 = 2.5\%.$$

It represents the total proportion of diseased individuals at the end of the year.

Step 3 — Display of Results

```
cat("Initial prevalence =", round(prevalence_init*100,2), "%\n")
cat("Annual incidence =", round(incidence*100,2), "%\n")
cat("Final prevalence =", round(prevalence_final*100,2), "%\n")
```

Explanation:

- `cat()` prints text to the console.
- `round(..., 2)` rounds results to two decimal places.
- Multiplication by 100 converts proportions into percentages.

Expected output:

```
Initial prevalence = 2 %
Annual incidence = 0.51 %
Final prevalence = 2.5 %
```

Step 4 — Data Frame for Visualization

```
data <- data.frame(
  Time = c("Jan 1", "Dec 31"),
  Prevalence = c(prevalence_init, prevalence_final) * 100
)
```

Explanation:

- A data frame is created to summarize the evolution of prevalence.
- The two time points (`"Jan 1"` and `"Dec 31"`) correspond to the beginning and end of the year.
- The prevalence values are multiplied by 100 to express them as percentages.

Step 5 — Visualization with `ggplot2`

```
library(ggplot2)

ggplot(data, aes(x = Time, y = Prevalence, group = 1)) +
  geom_line(color = "steelblue", linewidth = 1.5) +
  geom_point(size = 3, color = "darkred") +
  geom_text(aes(label = paste0(round(Prevalence,2), "%")),
            vjust = -1, color = "black", size = 4) +
  labs(
    title = "Change in Diabetes Prevalence Over the Year",
    x = "Time",
    y = "Prevalence (%)"
  ) +
  ylim(0, max(data$Prevalence)*1.3) +
  theme_minimal(base_size = 13)
```

Explanation:

- `ggplot(...)` initializes the graph.
- `geom_line()` draws the blue line showing prevalence change.
- `geom_point()` adds red points at the beginning and end of the year.
- `geom_text()` adds the numeric values above each point.
- `labs()` specifies the title and axis labels.
- `theme_minimal()` provides a clean visual style.

Graphical Representation

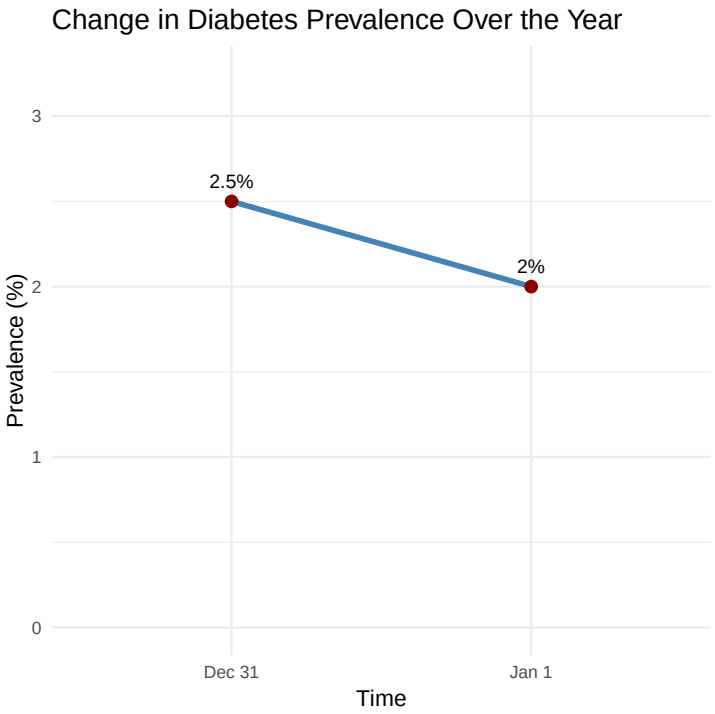


Figure 1.1: Evolution of diabetes prevalence between January and December. The plot shows a small increase in prevalence from 2% to 2.5%, reflecting the incidence over the year.

Statistical Interpretation

The results show that:

- prevalence rises from 2% to 2.5% in one year,
- annual incidence is 5 new cases per 1,000 inhabitants,
- the disease burden increases gradually over time.

Note. The R code presented in this chapter provides a ***detailed review*** *of the main epidemiological indicators: prevalence, incidence, and mortality. Each step is explained thoroughly, from defining the data to graphical visualization, allowing a clear understanding of the link between calculations and their statistical interpretation. These scripts serve as a* ***solid foundation*** *for the following chapters, where more advanced analyses will be introduced.*

1.6 Measuring Mortality

Mortality is a fundamental indicator of population health and a key measure for evaluating the effectiveness of health systems and interventions.

- **Crude death rate (CDR):**

$$CDR = \frac{\text{Number of deaths during a year}}{\text{Average population}} \times 1000$$

  It provides an overall picture of mortality but does not account for age structure.

- **Infant mortality rate:** number of deaths before age one per 1,000 live births.
- **Maternal mortality rate:** number of deaths of women related to pregnancy per 100,000 live births.

Indicator	**Formula / Unit**	**Interpretation**
Birth rate	$\frac{\text{Births}}{\text{Population}} \times 1000$	Demographic vitality
Prevalence	$\frac{\text{Existing cases}}{\text{Population}}$	Disease burden at a given time
Incidence	$\frac{\text{New cases}}{\text{At-risk population}}$	Frequency of new disease cases
Crude death rate	$\frac{\text{Deaths}}{\text{Population}} \times 1000$	Overall mortality
DALY	Years lost due to disease or disability	Global health impact
QALY	Life years weighted by quality	Health gain from interventions

Table 1.1: Summary of main health indicators.

Conclusion

Health indicators provide essential information for monitoring and evaluating the health status of populations. However, they must be interpreted carefully, taking into account demographic structures, data quality, and contextual factors. Standardization methods (such as age standardization) are often required to ensure fair and meaningful comparisons between populations or over time.

Exercise: Comparing Mortality

Problem Statement

Two cities (A and B) of equal size (100,000 inhabitants) report the following annual deaths:

- City A: 800 deaths.
- City B: 600 deaths.

1. Calculate the crude death rates.
2. Interpret.

Solution

1. **Crude death rates (CDR):**
   The crude death rate represents the number of deaths per year per 1,000 inhabitants:

$$CDR = \frac{\text{Number of deaths}}{\text{Total population}} \times 1000$$

   City A:

$$CDR_A = \frac{800}{100\,000} \times 1000 = 8\,‰$$

   City B:

$$CDR_B = \frac{600}{100\,000} \times 1000 = 6\,‰$$

2. **Interpretation:**
   City A has a higher crude death rate (8‰) than City B (6‰). However, this may reflect demographic differences rather than actual health disparities.

Practical Work 2: Comparing Mortality in R

Step 1 — Defining Population

```r
# Define population and deaths
population <- 100000
deaths_A <- 800
deaths_B <- 600
```

This section defines the basic information:

- `population` stores the total population (100,000 inhabitants per city).
- `deaths_A` and `deaths_B` store the total number of deaths recorded in each city.

These are the raw inputs required to compute mortality rates.

Step 2 — Calculating Crude Death Rates

```r
# Compute crude death rates (per 1000 inhabitants)
CDR_A <- (deaths_A / population) * 1000
CDR_B <- (deaths_B / population) * 1000

# Display the results
cat("CDR City A =", CDR_A, "\u2030\n")
cat("CDR City B =", CDR_B, "\u2030\n")
```

Here, the formula

$$CDR = \frac{\text{Deaths}}{\text{Population}} \times 1000$$

is implemented directly in R.

- The multiplication by 1000 expresses the rate *per thousand inhabitants*.
- The function `cat()` prints the results clearly in the R console with the "‰" (per thousand) symbol.

Step 3 — Creating a Graphical Comparison

```r
# Create a vector of rates and city labels
CDR <- c(CDR_A, CDR_B)
cities <- c("City A", "City B")

# Produce a bar plot
barplot(CDR, names.arg = cities, col = c("steelblue", "salmon"),
        ylim = c(0,10), main = "Crude Death Rates by City",
        ylab = "Deaths per 1,000 inhabitants")
```

This step visually compares the mortality rates between the two cities.

- `CDR` and `cities` store the computed values and labels.
- `barplot()` creates a bar chart with different colors for each city.
- `ylim = c(0,10)` ensures a consistent vertical scale.
- Titles and axis labels improve readability.

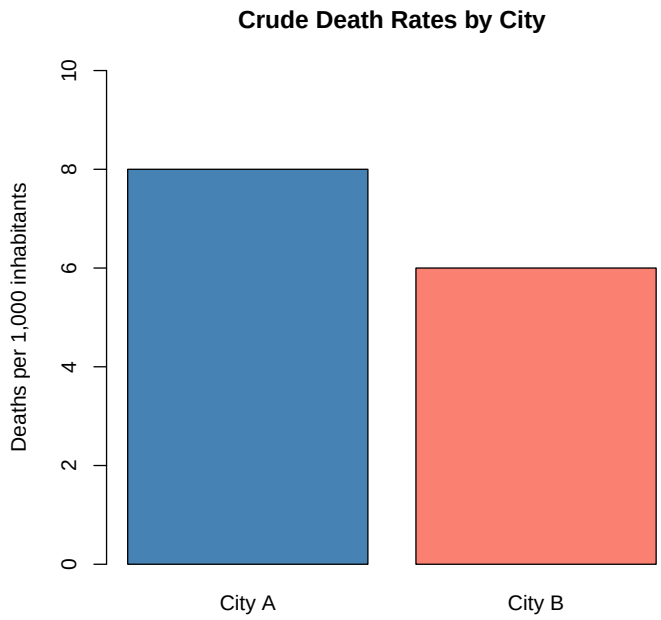


Figure 1.2: Comparison of crude death rates between City A and City B.

Step 4 — Statistical Interpretation

- City A shows a higher crude mortality rate than City B.
- This difference might reflect real health inequalities or simply demographic differences (e.g., older population in City A).
- **Conclusion:** To make valid comparisons, further analysis such as **age standardization** is recommended.

1.7 Chapter Conclusion

This chapter introduced the fundamentals of epidemiology and the main health indicators (prevalence, incidence, mortality). Understanding these notions is crucial for correctly analyzing population health and preparing for more advanced topics such as risk assessment, standardization, and regression in later chapters.

Transition Note. Chapter 1 introduced the fundamental epidemiological indicators used to describe population health. These measures provide essential descriptive insights but remain limited when studying complex relationships between variables. The next chapter therefore introduces statistical approaches for analyzing associations and risks in epidemiological data.

Chapter 2

Organization and Conduct of a Health Survey

2.1 Introduction

Health surveys constitute one of the principal empirical tools of epidemiology and public health. They allow the systematic collection, analysis, and interpretation of data describing the health status, behaviors, and living conditions of a defined population.

A health survey aims to:

- Quantify the occurrence and distribution of diseases (prevalence, incidence);
- Identify the determinants and risk factors (behavioral, environmental, socioeconomic);
- Evaluate the impact of public health programs or preventive actions;
- Guide decisions and resource allocation.

A poorly designed survey can produce biased results, threaten validity, and lead to erroneous inferences. Therefore, rigorous planning, methodological coherence, and ethical integrity are fundamental at each stage.

Reminder:

The study protocol is the backbone of any health survey. It should describe objectives, target population, study design, sampling strategy, instruments, and statistical analyses. This documentation ensures reproducibility and transparency.

2.2 Conceptual Foundations of Health Surveys

2.2.1 Defining the Objective and Research Question

The formulation of a clear, specific, and feasible research question guides the entire process. Let Q denote the central research question; it determines:

- The **study design** (cross-sectional, cohort, case–control, randomized controlled trial (RCT));

- The **variables** to be measured and their operational definitions;
- The **target population** and sampling frame;
- The **statistical analysis plan**.

The PICO structure is often adopted:

$$P = \text{Population}, \quad I = \text{Intervention or Exposure}, \quad C = \text{Comparison}, \quad O = \text{Outcome}.$$

Example. Research question: "What is the prevalence of hypertension among adults aged 30–70 years in Algiers?"

- P: adults aged 30–70 years residing in Algiers;
- I: not applicable (descriptive objective);
- C: none;
- O: presence of hypertension (binary variable).

2.2.2 Key Definitions and Notations

Health Indicator (Y): A measurable variable reflecting a health dimension (e.g., blood pressure, BMI, smoking status). Indicators may be quantitative (continuous, discrete) or qualitative (binary, ordinal, nominal).

Unit of Observation (i): The elementary entity providing data, such as an individual, household, or institution. Each unit contributes an observation $(Y_i, X_{i1}, \ldots, X_{ip})$.

Target Population (U): The full set of units of interest, denoted $U = \{1, 2, \ldots, N\}$. The aim of the study is to infer characteristics of U.

Sample (s): A subset of U, $s = \{i_1, i_2, \ldots, i_n\}$, selected according to a probability mechanism with inclusion probabilities $\pi_i = P(i \in s)$.

Sampling Weight (w_i): The inverse of the inclusion probability, $w_i = 1/\pi_i$, used to produce unbiased estimators of population quantities.

Example:

In a health survey involving a target population of $N = 10{,}000$ adults, a simple random sample of size $n = 1{,}000$ is selected. Each sampled unit has an inclusion probability

$$\pi_i = \frac{n}{N} = \frac{1{,}000}{10{,}000} = 0.1,$$

and thus a sampling weight

$$w_i = \frac{1}{\pi_i} = 10.$$

Each respondent therefore represents 10 individuals in the population.

The sample mean is given by

$$\bar{Y}_s = \frac{1}{n} \sum_{i \in s} Y_i,$$

which serves as an unbiased estimator of the population mean

$$\bar{Y} = \frac{1}{N} \sum_{i \in U} Y_i.$$

Review of Existing Evidence

Before data collection, a review of literature and official statistics enables:

- Estimation of expected parameters (e.g., $p_0 = 0.25$ for hypertension prevalence);
- Identification of validated measurement instruments (questionnaires, clinical measures);
- Anticipation of logistical, ethical, or cultural constraints;
- Alignment with previous surveys to ensure comparability.

Sources of Evidence.

- Scientific databases (PubMed, Scopus);
- National health surveys (DHS, WHO STEPS, national statistical offices);
- Health program reports, demographic yearbooks, and local studies.

2.2.3 Defining the Study Population

The **target population** is the entire set of units to which inferences will apply. The **sampling frame** is the list or operational definition of U that allows the drawing of a sample.

Definition 2.1 (Inclusion and Exclusion Criteria)**.** *These are explicit rules determining who can or cannot participate in the study:*

- ***Inclusion:*** *criteria that must be satisfied (e.g., age 30–70 years, residence* ≥ 6 *months, consent obtained);*
- ***Exclusion:*** *conditions preventing participation (e.g., severe illness, institutionalized individuals, communication incapacity).*

Example. For a study on diabetes risk among adults in a city:

- *Inclusion:* residents aged ≥ 18 years;
- *Exclusion:* pregnant women and individuals unable to respond.

2.2.4 Conceptual Model of Variables

A health survey often relies on a conceptual framework linking:

- **Dependent variable (Y):** health outcome (e.g., presence of disease);
- **Independent variables ($X_1, X_2, \ldots, X_p$):** determinants (socioeconomic status, lifestyle, biological factors);
- **Confounders (Z_j):** variables influencing both exposure and outcome.

Example: Confounding Effect of Age in the Relationship Between Smoking and Heart Disease

To illustrate the concept of a **confounder**, consider the following example involving three variables:

- X: **Smoking status** (1 = smoker, 0 = non-smoker);
- Y: **Heart disease occurrence** (1 = diseased, 0 = healthy);
- Z: **Age** (in years), a potential confounding variable.

Conceptual Explanation. We wish to study the effect of smoking (X) on the probability of having heart disease (Y). However, **age (Z)** may confound this relationship because:

1. Older individuals are more likely to smoke ($Z \to X$);
2. The risk of heart disease also increases with age ($Z \to Y$).

Thus, Z influences both the exposure (X) and the outcome (Y), creating a spurious association if not properly controlled.

Mathematical Representation. Ignoring the confounder leads to a simple (and biased) model:

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i,$$

where the coefficient α_1 captures not only the true effect of smoking but also part of the effect of age.

To obtain an unbiased estimate of the smoking effect, we must include age in the model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i,$$

where:

- β_1 represents the *net effect* of smoking on heart disease, adjusted for age;
- β_2 measures the effect of age on heart disease;
- ε_i is a random error term.

Interpretation. The inclusion of Z removes the distortion caused by the confounding pathway, it ensures that β_1 reflects the direct relationship between smoking and heart disease.

2.2.5 From Concept to Measurement

Operationalization involves transforming abstract constructs (e.g., "health perception") into measurable variables:

- **Objective measures:** blood pressure, weight, laboratory tests;
- **Subjective measures:** self-reported health, perceived stress;

- **Composite indices:** quality of life scales, socioeconomic indices.

Tip: Use standardized instruments whenever possible (e.g., WHOQOL, PHQ-9, IPAQ). This ensures reliability, comparability, and interpretability of results.

2.2.6 Ethical and Practical Considerations

Ethical principles (autonomy, beneficence, justice) must guide survey implementation:

- Informed consent obtained in writing;
- Data confidentiality and anonymization;
- Institutional ethics approval prior to fieldwork;
- Cultural sensitivity and community engagement.

Example of Ethical Statement. "The study protocol was approved by the Institutional Review Board of the University of Algiers. All participants provided written informed consent before participation."

Summary. Conceptual preparation is the cornerstone of survey validity. Defining clear objectives, understanding existing evidence, specifying the target population, and identifying variables of interest constitute the logical foundation upon which all subsequent methodological choices depend — including sampling, questionnaire design, and statistical analysis.

2.2.7 Estimation of Population Prevalence under Simple Random Sampling

In many health surveys, the objective is to estimate the prevalence of a given condition in a defined population. Let us denote by:

$$Y_i = \begin{cases} 1, & \text{if individual } i \text{ has the condition (e.g., hypertension);} \\ 0, & \text{otherwise.} \end{cases}$$

Each Y_i is a binary random variable defined for all individuals in the target population $U = \{1, 2, \ldots, N\}$.

Population parameter. The true prevalence (population proportion) is defined as:

$$p = \frac{1}{N} \sum_{i \in U} Y_i.$$

This value p is fixed but unknown in practice.

Sampling. Suppose a simple random sample s of size n is drawn from the population. Each unit has the same inclusion probability:

$$\pi_i = \frac{n}{N}, \quad \text{so that} \quad w_i = \frac{1}{\pi_i} = \frac{N}{n}.$$

Estimator. Under simple random sampling, the sample estimate of prevalence is:

$$\hat{p} = \frac{1}{n} \sum_{i \in s} Y_i,$$

where $Y_i = 1$ if individual i is affected and 0 otherwise. This estimator $\hat{p}$ is unbiased for the true population proportion p.

Remark 2.2.1. *The quantity p represents the* theoretical (population) prevalence*, that is, the true but unknown proportion of individuals affected by the disease in the entire target population. Because this parameter cannot be directly observed, health surveys estimate it empirically through the* sample prevalence $\hat{p}$ *computed from the collected data. Thus, $\hat{p}$ serves as an unbiased statistical estimate of p.*

Exercise: Estimation of Hypertension Prevalence

Problem Statement

Consider a finite population of $N = 10{,}000$ adults aged 30–70 years in Algiers. Suppose the true (but unknown) prevalence of hypertension is $p = 0.25$.

A simple random sample of size $n = 1{,}000$ is selected.

Answer the following questions:

1. Compute the inclusion probability π_i and the sampling weight w_i.
2. Write the mathematical expression of the estimator of the population prevalence under simple random sampling.
3. Suppose that 260 individuals in the sample are hypertensive. Compute the sample prevalence $\hat{p}$.
4. Provide an interpretation of the result at the population level.

Solution

1. **Inclusion Probability and Sampling Weight**

   Under simple random sampling:

   $$\pi_i = \frac{n}{N} = \frac{1{,}000}{10{,}000} = 0.1.$$

   The sampling weight is:

   $$w_i = \frac{1}{\pi_i} = \frac{1}{0.1} = 10.$$

   Thus, each sampled individual represents 10 individuals in the population.

2. **Estimator of the Population Prevalence**

   Let

   $$Y_i = \begin{cases} 1, & \text{if individual } i \text{ is hypertensive,} \\ 0, & \text{otherwise.} \end{cases}$$

The estimator of the population prevalence is:

$$\hat{p} = \frac{1}{n} \sum_{i \in s} Y_i.$$

This estimator is unbiased for the true population proportion p.

3. **Computation of the Sample Prevalence**

   If 260 individuals out of 1,000 are hypertensive:

$$\hat{p} = \frac{260}{1{,}000} = 0.26.$$

   Thus, the estimated prevalence is 26%.

4. **Interpretation**

   The sample prevalence $\hat{p} = 0.26$ suggests that approximately 26% of adults aged 30–70 years in Algiers are hypertensive.

   Since each sampled individual represents 10 individuals in the population, this corresponds to:

$$0.26 \times 10{,}000 = 2{,}600$$

   individuals in the target population.

   Although the true prevalence is unknown in practice, this example illustrates how the estimator operates under a simple random sampling design.

Practical Work 3: Estimation of Hypertension Prevalence in R

Listing 2.1: Estimation of a population proportion (hypertension example)

```r
################################################################
# Example: Estimation of a Population Proportion
# Objective: illustrate the estimation of hypertension prevalence
################################################################

# 1. Define the target population
N <- 10000                # total population size
p_true <- 0.25            # true (but unknown) prevalence in the
   population

# 2. Create the population variable Y
#    1 = hypertensive, 0 = not hypertensive
Y <- c(rep(1, N * p_true), rep(0, N * (1 - p_true)))

# 3. Draw a simple random sample of n = 1000 individuals
set.seed(2025)            # for reproducibility
n <- 1000
sample_index <- sample(1:N, n, replace = FALSE)
```

```r
Y_sample <- Y[sample_index]

# 4. (OPTION 1) Compute the actual sample prevalence from the random draw
prevalence_hat_random <- mean(Y_sample)

# 5. (OPTION 2) Force a deterministic result matching the theoretical example
#    Suppose exactly 260 individuals in the sample have Y = 1
Y_sample_fixed <- c(rep(1, 260), rep(0, n - 260))
prevalence_hat_fixed <- mean(Y_sample_fixed)

# 6. Display results
cat("Population size (N) =", N, "\n")
cat("Sample size (n)      =", n, "\n")
cat("True prevalence (p)  =", p_true, "\n\n")

cat("Random sample estimate (may vary) =", round(prevalence_hat_random, 3), "\n")
cat("Fixed theoretical example:         =", prevalence_hat_fixed, "\n")
```

Interpretation. This R script simulates a finite population of $N = 10{,}000$ individuals, where the true (but unknown) prevalence of hypertension is $p = 0.25$. Each individual is represented by a binary variable Y_i:

$$Y_i = \begin{cases} 1, & \text{if hypertensive;} \\ 0, & \text{otherwise.} \end{cases}$$

A simple random sample of $n = 1{,}000$ individuals is selected. The sample prevalence is estimated as:

$$\hat{p} = \frac{1}{n} \sum_{i \in s} Y_i.$$

When drawing a random sample, the result may vary slightly from the true value p. For example, the random draw might yield $\hat{p} = 0.23$, while a fixed illustrative case with 260 hypertensive individuals gives $\hat{p} = 0.26$.

The estimator $\hat{p}$ is unbiased, meaning that on average across repeated samples, its expected value equals the true population proportion p. Differences between $\hat{p}$ and p reflect only **sampling variability**.

2.3 Choice of Study Design

The study design determines the logic by which observations are collected and causal or associative relationships are inferred. It must align with the research question, available resources, and the desired level of evidence.

Definition 2.2. *A study design specifies the temporal structure of data collection, the exposure–outcome relationship, and the mechanism of participant selection.*

Four major designs are commonly employed in health research: cross-sectional, cohort, case–control, and randomized controlled trial (RCT). Each has distinct objectives, strengths, and limitations.

2.3.1 Cross-Sectional Surveys

A **cross-sectional design** collects information at a single point in time. It provides a "snapshot" of the population, allowing estimation of disease **prevalence** and exploration of associations between outcomes and risk factors.

Statistical Principle. Let Y_i be a binary indicator (e.g., disease presence) for subject i, $i = 1, \ldots, n$. The sample *prevalence estimator* is:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

with variance under simple random sampling (SRS):

$$Var(\hat{p}) = \frac{p(1-p)}{n} \times \frac{N-n}{N-1},$$

where N is the finite population size. The finite population correction factor $\frac{N-n}{N-1}$ accounts for sampling without replacement.

Confidence Interval for a Proportion. Under simple random sampling, the sample proportion $\hat{p}$ is an unbiased estimator of the population prevalence p ($\mathbb{E}[\hat{p}] = p$). By the Central Limit Theorem (CLT), when n is large, the sampling distribution of $\hat{p}$ is approximately normal:

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

Thus, a $(1-\alpha) \times 100\%$ confidence interval for p is given by:

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

where $z_{1-\alpha/2}$ is the critical value from the standard normal distribution (e.g., $z_{0.975} = 1.96$ for a 95% CI). If the sampling fraction n/N is not negligible, the variance can be adjusted using the finite population correction factor $\frac{N-n}{N-1}$.

Example 1: Prevalence in a Cross-Sectional Survey

Statement. In a cross-sectional survey of $n = 1000$ adults, 250 report hypertension ($Y_i = 1$). We want to estimate the **prevalence** of hypertension.

Data

- Sample size: $n = 1000$
- Number with condition: $X = 250$

Point Estimate

$$\hat{p} = \frac{X}{n} = \frac{250}{1000} = 0.25$$

Standard Error (SRS, no FPC)

Estimate the variance by plugging $\hat{p}$ into $Var(\hat{p}) = \frac{p(1-p)}{n}$:

$$Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.25 \times 0.75}{1000} = 0.0001875$$

$$SE(\hat{p}) = \sqrt{0.0001875} \approx 0.0136931$$

95% Confidence Interval (Normal Approximation)

Using $z_{0.975} = 1.96$:

$$\hat{p} \pm 1.96 \times SE(\hat{p}) = 0.25 \pm 1.96 \times 0.0136931 = 0.25 \pm 0.0268384$$

Thus:

$$95\%\ \text{CI} = [0.22316,\ 0.27684] \approx [22.3\%,\ 27.7\%]$$

Conclusion

Estimated prevalence: $\hat{p} = 0.25$ (25%).
95% CI (no FPC): $[0.22316,\ 0.27684] \approx [22.3\%,\ 27.7\%]$

Interpretation of the Finite Population Correction (FPC)

The factor

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

reduces the variance when sampling *without replacement* from a finite population. If the sample is a small fraction of the population (e.g., $n/N < 0.05$), the correction factor is approximately 1 and can be ignored; otherwise, it should be included.

Example 2: Factors Associated with Obesity in a Cross-Sectional Survey

Statement. Researchers aimed to investigate the **factors associated with obesity** among adults in a specific city in 2025. A random sample of 1,000 individuals was selected, and each participant completed a questionnaire and underwent measurements of weight and height.

Variables of Interest

The analysis explored the relationship between obesity and the following factors:

- **Age:** Does obesity increase with advancing age?
- **Dietary habits:** Are individuals who frequently consume fast food more likely to be obese?

- **Smoking status:** Are smokers less or more likely to be obese?
- **Physical activity:** Is low physical activity associated with higher body mass index (BMI)?

Results

- Overall prevalence of obesity: 28%
- Obesity was associated with:
  - Low physical activity
  - Frequent consumption of fast food

Interpretation

Since the data were collected at a *single point in time*, causality cannot be inferred. While fast-food consumption and obesity are correlated, one cannot conclude that fast food *causes* obesity.

Advantages.

- Simple, quick, and cost-efficient.
- Suitable for estimating prevalence of diseases and risk factors.
- Enables simultaneous study of multiple variables.

Limitations.

- Temporal ambiguity: exposure and outcome measured simultaneously.
- Cannot establish causality.
- Sensitive to survival bias (long-duration cases overrepresented).

Interpretation: Cross-sectional results describe "what exists" rather than "what causes what."

2.3.2 Cohort Studies

A **cohort study** follows individuals over time to estimate the incidence or risk of developing a health event. Participants are classified by exposure status at baseline and observed for outcome occurrence.

Notation and Estimation. Let T_i denote follow-up time and δ_i an event indicator ($\delta_i = 1$ if event occurs, 0 otherwise). The incidence rate estimator is:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i}.$$

If participants are grouped by exposure ($E = 1$ vs. $E = 0$), the relative risk is:

$$RR = \frac{\hat{\lambda}_1}{\hat{\lambda}_0}.$$

Example 3: Incidence and Relative Risk in a Cohort Study

Statement. A cohort study followed 1000 non-smokers and 1000 smokers for 5 years to investigate the incidence of cardiovascular disease. We want to estimate the incidence rates in each group and the relative risk (RR) of cardiovascular disease associated with smoking.

Data

- Non-smokers: $n_0 = 1000$, number of cases $X_0 = 50$
- Smokers: $n_1 = 1000$, number of cases $X_1 = 150$
- Follow-up time per person: 5 years

Point Estimates

$$\hat{\lambda}_0 = \frac{X_0}{n_0 \times 5} = \frac{50}{5000} = 0.01$$

$$\hat{\lambda}_1 = \frac{X_1}{n_1 \times 5} = \frac{150}{5000} = 0.03$$

$$RR = \frac{\hat{\lambda}_1}{\hat{\lambda}_0} = \frac{0.03}{0.01} = 3.0$$

Interpretation

The incidence rate of cardiovascular disease is three times higher in smokers compared to non-smokers. This cohort study demonstrates a temporal relationship where exposure (smoking) precedes the outcome (cardiovascular disease), allowing for estimation of relative risk.

Advantages.

- Establishes temporal order (exposure precedes outcome).
- Allows estimation of incidence and relative risk.
- Can study multiple outcomes for a single exposure.

Limitations.

- Requires long follow-up and substantial resources.
- Attrition (loss to follow-up) may bias results.
- Inefficient for rare diseases.

Tip: Cohort studies may be *prospective* (data collected forward in time) or *retrospective* (based on existing records).

2.3.3 Case–Control Studies

A **case–control design** compares individuals with a disease (cases) to those without (controls) to investigate past exposures. This design is retrospective and particularly useful for rare diseases.

Analytical Framework. Data are summarized in a 2×2 contingency table:

	Exposed	Unexposed	Total
Cases	a	b	$a+b$
Controls	c	d	$c+d$
Total	$a+c$	$b+d$	n

The measure of association is the **odds ratio (OR)**:

$$OR = \frac{ad}{bc}.$$

If OR $= 1 \rightarrow$ no association.

If OR $> 1 \rightarrow$ positive association (the exposure increases the risk of disease).

If OR $< 1 \rightarrow$ protective association (the exposure decreases the risk of disease).

Estimation of Precision. The standard error of the logarithm of the odds ratio is:

$$SE_{\log(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

A 95% confidence interval on the log scale is:

$$\log(OR) \pm 1.96 \times SE_{\log(OR)}.$$

Exponentiating gives the confidence interval on the OR scale:

$$CI = \left(e^{\log(OR) - 1.96\, SE_{\log(OR)}},\ e^{\log(OR) + 1.96\, SE_{\log(OR)}} \right).$$

Remark 2.3.1 (Why Take the Logarithm and Compute the Variance of the Odds Ratio)**.** *The odds ratio (OR) is defined as:*

$$OR = \frac{ad}{bc}.$$

It represents a ratio of ratios and can take values from 0 *to* $+\infty$*. However, this scale is highly asymmetric and inconvenient for statistical inference.*

- *Symmetry and Normality. The logarithm transforms the OR to a more symmetric and analytically convenient scale:*

$$\log(OR) = \log(a) + \log(d) - \log(b) - \log(c).$$

On this log scale:

  - $\log(OR) = 0$ *corresponds to* $OR = 1$ *(no association),*
  - $\log(OR) > 0$ *indicates a positive association,*

– $\log(OR) < 0$ *indicates a negative or protective association.*

This transformation makes the distribution of $\log(OR)$ *approximately normal, especially in large samples, which simplifies statistical inference.*

- *Simplification of Variance Calculations. Because the logarithm converts products and quotients into sums and differences, the variance of* $\log(OR)$ *becomes the sum of variances of its logarithmic components:*

$$Var[\log(OR)] = Var[\log(a)] + Var[\log(d)] + Var[\log(b)] + Var[\log(c)],$$

assuming the cell counts a, b, c, d *are independent.*

Delta Method Approximation. Using the *Delta Method*, a common statistical approximation for the variance of nonlinear functions, we have:

$$Var[\log(a)] \approx \frac{1}{a}, \quad Var[\log(b)] \approx \frac{1}{b}, \quad Var[\log(c)] \approx \frac{1}{c}, \quad Var[\log(d)] \approx \frac{1}{d}.$$

Substituting these gives the well-known result:

$$Var[\log(OR)] \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d},$$

and therefore the standard error is:

$$SE_{\log(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

Practical Implications. Since the confidence intervals are computed on the log scale:

$$\log(OR) \pm 1.96 \times SE_{\log(OR)},$$

we exponentiate the limits to obtain the interval on the original OR scale:

$$CI = \left(e^{\log(OR) - 1.96\, SE_{\log(OR)}},\ e^{\log(OR) + 1.96\, SE_{\log(OR)}}\right).$$

Working on the log scale ensures additive effects, symmetry around the null value ($OR = 1$), and easier interpretation.

Conclusion: Taking the logarithm of the odds ratio is a standard and essential step that simplifies inference, stabilizes variance, and provides a more interpretable, symmetric scale for analysis.

Example 4: Association Between Smoking and Lung Cancer in a Case–Control Study

Statement. The objective is to investigate whether smoking is associated with lung cancer.

Study Design

A total of 400 participants were selected:

- 200 individuals with lung cancer (*cases*)
- 200 individuals without lung cancer (*controls*)

Both groups were asked about their past smoking habits.

	Smokers (Exposed)	Non-smokers (Unexposed)	Total
Cases (disease)	120	80	200
Controls (no disease)	60	140	200
Total	180	220	400

Data Summary

Calculation of the Odds Ratio (OR)

$$OR = \frac{a \times d}{b \times c} = \frac{120 \times 140}{80 \times 60} = 3.5$$

$$SE_{\log(OR)} = \sqrt{\frac{1}{120} + \frac{1}{80} + \frac{1}{60} + \frac{1}{140}} \approx 0.211$$

95% Confidence Interval:

$$\log(3.5) \pm 1.96 \cdot 0.211 \Rightarrow (0.838, 1.668)$$

$$CI = (e^{0.838}, e^{1.668}) \approx (2.31, 5.30)$$

Interpretation

There is a strong association between smoking and lung cancer, as indicated by the odds ratio $OR = 3.5$ and the 95% CI $(2.31, 5.30)$.

Advantages.

- Efficient for rare outcomes or long-latency diseases.
- Requires fewer subjects and resources.
- Allows study of multiple exposures for one disease.

Limitations.

- Susceptible to recall bias (inaccurate memory of exposures).
- Selection bias possible in control recruitment.
- Cannot estimate incidence directly.

Remark 2.3.2. ***Interpretation:*** *When disease is rare ($p < 0.1$), the odds ratio approximates the relative risk: $OR \approx RR$.*

Intuitive Idea

The **Relative Risk (RR)** measures how many times the risk of disease increases among the exposed compared to the unexposed group:

$$RR = \frac{Risk_{exposed}}{Risk_{unexposed}} = \frac{p_1}{p_0}$$

The **Odds Ratio (OR)** measures how many times the odds of disease increase among the exposed compared to the unexposed group:

$$OR = \frac{odds_{exposed}}{odds_{unexposed}} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

When the disease is rare (that is, when both p_1 and p_0 are very small), the terms $(1-p_1)$ and $(1-p_0)$ are approximately equal to 1. Hence:

$$OR \approx \frac{p_1}{p_0} = RR$$

Therefore, for rare diseases, the Odds Ratio closely approximates the Relative Risk.

2.3.4 Randomized Controlled Trials (RCTs)

A **randomized controlled trial (RCT)** evaluates the efficacy of an intervention under controlled conditions. Participants are randomly assigned to a treatment group ($I = 1$) or a control group ($I = 0$).

Principle. Randomization ensures, under large n, that both known and unknown confounders are balanced between groups. Thus, any observed difference in outcomes can be causally attributed to the intervention.

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$$

Variance and Standard Error.

$$Var(\hat{\alpha}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}, \quad SE(\hat{\alpha}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$$

Confidence Interval. A 95% confidence interval for the mean difference is:

$$\hat{\alpha} \pm 1.96 \times SE(\hat{\alpha})$$

Hypothesis Testing. To assess whether the treatment has a significant effect, we use the t-statistic:

$$t = \frac{\hat{\alpha}}{SE(\hat{\alpha})}$$

where the hypotheses are defined as:

$$H_0 : \alpha = 0 \quad \text{(no treatment effect)}, \qquad H_1 : \alpha \neq 0 \quad \text{(presence of effect)}$$

Under H_0, the test statistic approximately follows a Student's t-distribution with

$$df = n_1 + n_0 - 2$$

p-value. The p-value represents the probability of observing a t value as extreme as (or more extreme than) the one obtained, under the null hypothesis:

$$p = 2 \times P(|T| \geq |t_{obs}|)$$

If $p < \alpha$ (e.g., 0.05), we reject H_0 and conclude that the intervention has a statistically significant effect.

Example 5: Effect of a New Antihypertensive Drug in a Randomized Controlled Trial

Statement. We are testing whether a new antihypertensive drug reduces mean systolic blood pressure.

Data

$$\bar{Y}_1 = 130, \quad \bar{Y}_0 = 138, \quad s_1 = s_0 = 10, \quad n_1 = n_0 = 100$$

Point Estimate

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0 = 130 - 138 = -8$$

Interpretation: On average, the treatment group has a mean systolic blood pressure 8 mmHg lower than the control group.

Standard Error

$$SE(\hat{\alpha}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}} = \sqrt{\frac{10^2}{100} + \frac{10^2}{100}} = \sqrt{2} = 1.4142$$

Test Statistic

$$t = \frac{\hat{\alpha}}{SE(\hat{\alpha})} = \frac{-8}{1.4142} = -5.657$$

Degrees of Freedom

Since both groups have equal variances and equal sample sizes:

$$df = n_1 + n_0 - 2 = 100 + 100 - 2 = 198$$

p-value

The p-value is obtained from the Student's t distribution with 198 degrees of freedom:

$$p = 2 \times P(T_{198} > |t|) = 2 \times P(T_{198} > 5.657)$$

Using statistical tables or software:

$$P(T_{198} > 5.657) = 2.67 \times 10^{-8}; \quad \text{so} \quad p = 2 \times 2.67 \times 10^{-8} = 5.34 \times 10^{-8} < 0.001$$

Final Results

$$\text{Mean difference} = -8, \quad SE = 1.41, \quad t = -5.66, \quad df = 198, \quad p = 5.34 \times 10^{-8}$$

Conclusion

Since $p < 0.001$, we reject the null hypothesis ($H_0 : \alpha = 0$). The new drug significantly lowers mean systolic blood pressure by 8 mmHg compared with the control group.

Advantages.

- Strongest evidence for causality.
- Minimizes confounding and selection bias.
- Facilitates precise estimation of treatment effects.

Limitations.

- Ethical and practical constraints (cannot assign harmful exposures).
- Costly and time-consuming.
- Limited generalizability (highly controlled settings).

Educational Note: The research question must dictate the design—not the reverse. The design defines how evidence will be built, not merely how data will be collected.

Conclusion.

The choice of design balances scientific rigor, feasibility, cost, and ethics. A good design maximizes internal validity while maintaining sufficient external validity for generalization.

Table 2.1: Key Differences Between Major Epidemiological Study Designs

Feature	**Cross-sectional**	**Cohort**	**Case–Control**	**Randomized Controlled Trial (RCT)**
Time Direction	One point in time	Forward (exposure → outcome)	Backward (outcome → exposure)	Forward (randomized exposure)
Main Measure	Prevalence	Incidence, Risk	Odds Ratio	Risk or Risk Difference
Purpose	Describe and explore associations	Identify causes and estimate risk	Identify risk factors for rare diseases	Test causal effects of interventions
Causality Strength	Weak	Moderate to strong	Moderate	Strongest
Advantages	Quick, low cost	Temporal sequence clear	Good for rare diseases	Minimizes bias, high internal validity
Limitations	No temporal order	Expensive, long follow-up	Recall and selection bias	Ethical and cost issues

Exercise: Comparing Study Designs

Objective. This exercise summarizes the previous examples presented in this section and synthesizes their key methodological and statistical elements. The table 2.2 summarizes the purpose, key calculations, and interpretation of four common epidemiological study designs: *cross-sectional*, *cohort*, *case–control*, and *randomized controlled trial (RCT)*. It provides a practical framework for implementation and interpretation in R.

Summary of Study Objectives and Key Results

Study Design	**Objective**	**Key Result**	**Interpretation**
Cross-Sectional	Estimate prevalence of hypertension	$\hat{p} = 0.25$ (95% CI: 0.223–0.277)	25% of adults have hypertension at this point in time.
Cohort	Compare incidence of CVD in smokers vs non-smokers	$\lambda_0 = 0.01$, $\lambda_1 = 0.03$, $RR = 3$	Smokers have three times higher risk of CVD over 5 years.
Case–Control	Assess association between smoking and lung cancer	$OR = 3.5$ (95% CI: 2.31–5.30)	Strong positive association; smokers are more likely to have lung cancer.
RCT	Evaluate effect of new drug on systolic BP	$\Delta = -8$ mmHg, $p < 0.001$	Drug significantly lowers BP compared to control.

Table 2.2: Summary of study objectives, key results, and interpretation. Columns have fixed width to ensure proper line wrapping and readability.

Practical Work 4: Comparing Study Designs in R

Listing 2.2: R Implementation for the Four Study Designs

```
1
2  # ============================================================
3  # Applied Biostatistics Exercise: Four Study Designs
4  # ============================================================
5
6  # ----------------------------
7  # Q1: Cross-Sectional Survey
8  # ----------------------------
9  n <- 1000           # sample size
10 cases <- 250        # number of individuals with hypertension
11
12 # Point estimate
13 p_hat <- cases / n
14 # Standard error
15 SE <- sqrt(p_hat * (1 - p_hat) / n)
```

```r
# 95% CI
CI <- p_hat + c(-1.96, 1.96) * SE

cat("=== Cross-Sectional Survey ===\n")
cat("Prevalence =", round(p_hat,3), "\n")
cat("95% CI:", round(CI,3), "\n\n")

# Bar plot
barplot(c(cases, n-cases), names.arg=c("Hypertension","No Hypertension"),col=c("red","green"), main="Cross-Sectional Survey", ylab="Count")


# ------------------------------
# Q2: Cohort Study
# ------------------------------
n0 <- 1000          # non-smokers
n1 <- 1000          # smokers
years <- 5

cases0 <- 50        # CVD in non-smokers
cases1 <- 150       # CVD in smokers

# Incidence rates
lambda0 <- cases0 / (n0 * years)
lambda1 <- cases1 / (n1 * years)

# Relative Risk
RR <- lambda1 / lambda0

cat("=== Cohort Study ===\n")
cat("Incidence non-smokers:", round(lambda0,3), "\n")
cat("Incidence smokers:", round(lambda1,3), "\n")
cat("Relative Risk (RR):", round(RR,2), "\n\n")

# Plot incidence comparison
barplot(c(lambda0, lambda1), names.arg=c("Non-smokers","Smokers"),
        col=c("blue","orange"), main="Cohort Study: Incidence Rate", ylab="Incidence per year")


# ------------------------------
# Q3: Case-Control Study
# ------------------------------
a <- 120
b <- 80
c <- 60
d <- 140

# Odds Ratio
```

```r
OR <- (a*d)/(b*c)
# Standard error of log(OR)
SE_logOR <- sqrt(1/a + 1/b + 1/c + 1/d)
# 95% CI
CI_log <- log(OR) + c(-1.96,1.96)*SE_logOR
CI_OR <- exp(CI_log)

cat("=== Case-Control Study ===\n")
cat("Odds Ratio (OR):", round(OR,2), "\n")
cat("95% CI:", round(CI_OR,2), "\n\n")

# Mosaic plot
matrix_CC <- matrix(c(a,b,c,d), nrow=2, byrow=TRUE)
colnames(matrix_CC) <- c("Exposed","Unexposed")
rownames(matrix_CC) <- c("Cases","Controls")
mosaicplot(matrix_CC, main="Case-Control Study", col=c("red","
   green"))


# ------------------------------
# Q4: Randomized Controlled Trial (RCT)
# ------------------------------
n1 <- 100
n0 <- 100
mean1 <- 130
mean0 <- 138
sd1 <- 10
sd0 <- 10

# Mean difference
alpha <- mean1 - mean0
# Standard error
SE_delta <- sqrt(sd1^2/n1 + sd0^2/n0)
# t-statistic
t_value <- delta / SE_delta
# p-value
df <- n1 + n0 - 2
p_value <- 2 * pt(abs(t_value), df=df, lower.tail = FALSE)

cat("=== Randomized Controlled Trial ===\n")
cat("Mean difference (Treatment - Control):", delta, "\n")
cat("t-value:", round(t_value,2), "\n")
cat("p-value:", format(p_value, scientific=TRUE), "\n\n")

# Boxplot for visual comparison
boxplot(rnorm(n1, mean=mean1, sd=sd1),
        rnorm(n0, mean=mean0, sd=sd0),
        names=c("Treatment","Control"),
        col=c("skyblue","pink"),
        main="RCT: Systolic Blood Pressure",
        ylab="Systolic BP (mmHg)")
```

Visual Summary of Study Results

1. Cross-Sectional Survey: Hypertension Prevalence

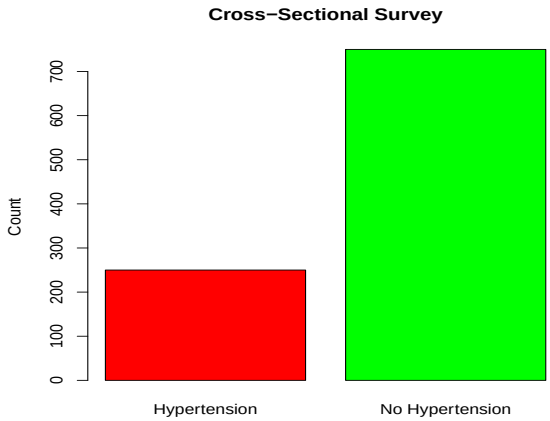


Figure 2.1: Distribution of hypertension in the sample (Cross-Sectional Survey). 25% of participants have hypertension.

Interpretation: This bar plot shows the proportion of adults with and without hypertension in the survey.
- 25% of participants have hypertension (250/1000).
- 75% of participants do not have hypertension (750/1000).
Note: This cross-sectional design estimates prevalence at a single point in time. It does not provide information about risk, incidence, or causality. It simply describes "what exists" in the population at the survey time.

2. Cohort Study: Incidence Rate of Cardiovascular Disease

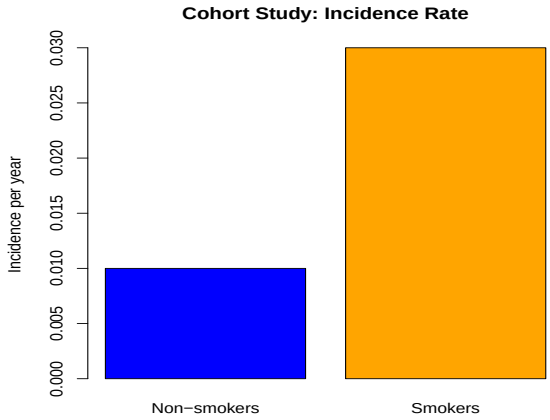


Figure 2.2: Incidence rate comparison between non-smokers and smokers.

Interpretation: Smokers show a threefold higher incidence rate of cardiovascular disease compared to non-smokers over 5 years. Cohort design allows estimation of risk over time.

3. Case–Control Study: Smoking and Lung Cancer

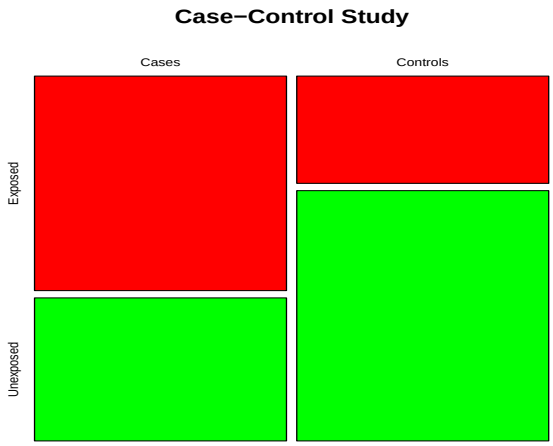


Figure 2.3: Mosaic plot showing the distribution of smoking among cases and controls.

Interpretation: The plot highlights the strong association between smoking and lung cancer. Odds of being a smoker are higher in cases than controls. This design estimates association, not incidence.

4. Randomized Controlled Trial: Blood Pressure

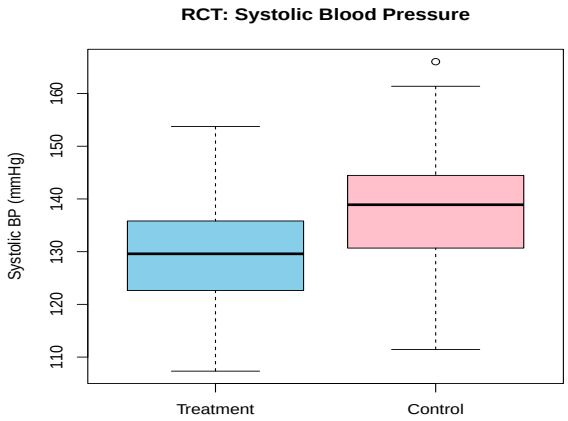


Figure 2.4: Boxplot of systolic blood pressure in treatment vs. control groups.

Interpretation: The treatment group shows a clear reduction in blood pressure compared with the control group. RCT allows causal inference about the drug's effect.

Overall Summary:

These plots provide a visual comparison of the four study designs: prevalence (cross-sectional), risk over time (cohort), association (case-control), and causal effect (RCT). They facilitate understanding of the objectives and interpretation of each design.

2.4 Chapter Conclusion

This chapter provided a comprehensive overview of the foundational concepts in epidemiology and health data analysis. It began by defining the main objectives of epidemiology and the role of health indicators in quantifying disease frequency, distribution, and determinants within populations. Various study designs were discussed, including cross-sectional, cohort, and case–control studies, each with its methodological characteristics, advantages, and potential sources of bias.

Special attention was given to the interpretation of key measures such as prevalence, incidence, risk ratios, and odds ratios, along with the importance of understanding confounding and bias in observational research. The chapter also emphasized the conceptual framework for study design through the PICO model, linking population, intervention or exposure, comparison, and outcome.

Mathematical formulations were presented to illustrate sampling principles, estimation of population parameters, and the use of weights to correct for unequal probabilities of selection. Additionally, applied examples using `R` code demonstrated how to compute descriptive statistics, visualize data, and fit simple epidemiological models.

In summary, this chapter established the methodological foundation necessary for conducting and interpreting applied epidemiological studies. It bridged theoretical understanding and computational practice.

Transition Note. Chapter 2 presented the foundations of a health survey: defining objectives, selecting the study design, specifying the target population, and operationalizing variables. Chapter 3 will focus on data collection, processing, and analysis techniques, showing how these concepts are translated into reliable and interpretable results for public health.

Chapter 3

Standardization Methods in Epidemiology

3.1 Introduction

In epidemiology, crude comparisons of morbidity or mortality rates between populations can be misleading when the populations differ in age structure or other demographic characteristics. Standardization is a statistical procedure used to remove the influence of these structural differences, allowing fair and meaningful comparisons.

Two main approaches exist:

- **Direct standardization**
- **Indirect standardization** (based on the Standardized Mortality Ratio, SMR)

This chapter explains both methods with definitions, formulas, practical examples, interpretations, and exercises, including R implementations.

3.2 Why Standardize?

Standardization is necessary when:

- Populations differ in age distribution,
- Crude rates may give misleading impressions of risk,
- Public health decisions require fair comparisons,
- Comparing regions, hospitals, or time periods.

3.3 Key Definitions

Crude Rate

Let D be the number of observed events and N the population size. The crude rate is

$$T_{\text{crude}} = \frac{D}{N} \times k,$$

where k is typically 1,000 or 100,000.

Examples:

- The overall mortality rate in a country:

$$T_{\text{crude}} = \frac{D}{N} \times 100{,}000$$

  where D is total deaths and N is the total population.

- The crude incidence rate of influenza in a city during one year.

Stratum-Specific Rate

For stratum i with D_i events and N_i individuals,

$$T_i = \frac{D_i}{N_i} \times k.$$

These are essential for both direct and indirect standardization.

Examples:

- Age-specific mortality rate for individuals aged 65–74:

$$T_{65-74} = \frac{D_{65-74}}{N_{65-74}} \times 100{,}000$$

- Sex-specific incidence rate of lung cancer (e.g., rate among males only).

Difference between Crude Rate and Stratum-Specific Rate

- **Crude Rate:** A single overall rate calculated for the entire population, without considering age or any other stratification.
- **Stratum-Specific Rate:** A rate calculated separately for each subgroup (stratum) of the population, such as age groups, sex, or regions, providing a more precise measure of risk within each category.

3.4 Direct Standardization

3.4.1 Concept

Direct standardization answers the question:

> "What would the rate be if the study population had the same structure as a reference (standard) population?"

It applies the **study population's stratum-specific rates** to a fixed reference distribution.

3.4.2 Formula

Let T_i be the study population's rate in stratum i, and w_i the size of stratum i in the reference population.

The directly standardized rate is:

$$T_{\text{std}}^{(\text{direct})} = \frac{\sum_i w_i T_i}{\sum_i w_i}.$$

It represents the hypothetical rate the **reference population** would experience if it were exposed to the **study population's risks**.

Example 01: Direct Standardization with Two Age Strata

Statement

We want to calculate the directly standardized rate using two age strata. The study population provides age-specific rates, and the standard population provides the age distribution.

Data

Age Group	**Study Population Rate T_i (per 100,000)**	**Standard Population Size w_i**
0–49	50	80,000
50+	400	20,000

Formula

The directly standardized rate is defined as:

$$T_{\text{std}}^{(\text{direct})} = \frac{\sum_i w_i T_i}{\sum_i w_i}$$

Step 1: Compute the Numerator

$$w_1 T_1 = 80{,}000 \times 50 = 4{,}000{,}000$$

$$w_2 T_2 = 20{,}000 \times 400 = 8{,}000{,}000$$

$$\sum_i w_i T_i = 4{,}000{,}000 + 8{,}000{,}000 = 12{,}000{,}000$$

Step 2: Compute the Denominator

$$\sum_i w_i = 80{,}000 + 20{,}000 = 100{,}000$$

Step 3: Compute the Directly Standardized Rate

$$T_{\text{std}}^{(\text{direct})} = \frac{12{,}000{,}000}{100{,}000} = 120 \text{ per } 100{,}000$$

Conclusion

If the standard population were exposed to the study population's age-specific rates, it would experience 120 cases per 100,000.

Example 02: Direct Standardization Comparing Two Cities

Statement

Consider two cities (A and B) with different age distributions. We use a common standard population to compute directly standardized rates in order to compare mortality fairly.

Data

Age Group	Standard Pop. w_i	Rate A ($T_{i,A}$)	Rate B ($T_{i,B}$)
0–44	50,000	2	4
45–64	30,000	10	12
65+	20,000	30	35

Formula

$$T_{\text{std}}^{(\text{direct})} = \frac{\sum_i w_i T_i}{\sum_i w_i}$$

Step 1: Directly Standardized Rate for City A

$$T_{\text{std},A} = \frac{50{,}000 \times 2 + 30{,}000 \times 10 + 20{,}000 \times 30}{50{,}000 + 30{,}000 + 20{,}000} = \frac{1{,}000{,}000}{100{,}000} = 10$$

Step 2: Directly Standardized Rate for City B

$$T_{\text{std},B} = \frac{50{,}000 \times 4 + 30{,}000 \times 12 + 20{,}000 \times 35}{100{,}000} = \frac{1{,}260{,}000}{100{,}000} = 12.6$$

Conclusion

Directly standardized rates adjust for differences in age distribution. City B has a higher risk-adjusted rate (12.6) than City A (10), indicating higher mortality after age adjustment.

3.4.3 Advantages and Limitations

Advantages

- **Allows fair comparison between populations:** adjusts for differences in age (or other strata) and removes the influence of population structure.
- **Uses the study population's actual rates:** comparisons reflect the observed risks in each group.

- **Easy to interpret:** answers the hypothetical question "What would the rate be if both populations had the same age distribution?"
- **Useful for comparing multiple populations:** as long as the same standard population is used, comparisons remain consistent.
- **More stable with large populations:** large strata produce more reliable stratum-specific rates.

Limitations

- **Requires stable stratum-specific rates:** rare outcomes or small strata produce unstable rates and unreliable standardized results.
- **Needs detailed data:** stratum-specific rates must be available for each population; crude-only data are insufficient.
- **Dependent on the choice of standard population:** different standards may produce different standardized rates.
- **Produces a hypothetical measure:** the standardized rate is not a real rate but a constructed one for comparison only.
- **Partial adjustment:** standardization only adjusts for the chosen stratification variable; other confounders may remain.

3.5 Indirect Standardization

3.5.1 Concept

Indirect standardization answers the question:

> "How many events would occur in the study population if it experienced the reference population's stratum-specific rates?"

This method is particularly useful when the study population's stratum-specific rates are unstable or unreliable (e.g., small populations).

3.5.2 Formula

Suppose:

- N_i = size of the study population in stratum i,
- T_i^{ref} = reference rate in stratum i,
- k = rate multiplier (e.g., $k = 100$, 1000, or 100,000),
- O = total observed number of events in the study population.

If the reference rates are expressed per k individuals, they must first be converted into probabilities:

$$\lambda_i^{\text{ref}} = \frac{T_i^{\text{ref}}}{k}.$$

Expected Number of Events

The expected number of events is:

$$E = \sum_i N_i \lambda_i^{\text{ref}} = \sum_i N_i \frac{T_i^{\text{ref}}}{k}.$$

Standardized Mortality Ratio (SMR)

$$\text{SMR} = \frac{O}{E}.$$

Interpretation:

- SMR $= 1$: observed equals expected.
- SMR > 1: excess risk compared to reference.
- SMR < 1: lower-than-expected risk.

Example 1: Mortality in a small hospital (Rates per 100)

Statement

A small hospital has 200 patients distributed across three age groups. The reference population provides age-specific mortality rates expressed per 100 people. We want to compute the expected number of deaths and the Standardized Mortality Ratio (SMR).

Data

- Total patients: 200

Age Group	N_i	T_i^{ref} (per 100)
0–49	100	0.2
50–69	70	1
70+	30	5

Step 1: Convert Rates

$$\lambda_i^{\text{ref}} = \frac{T_i^{\text{ref}}}{100}$$

$$\lambda_{0-49} = 0.002, \quad \lambda_{50-69} = 0.01, \quad \lambda_{70+} = 0.05.$$

Step 2: Compute Expected Deaths

$$E = 100 \times 0.002 + 70 \times 0.01 + 30 \times 0.05$$

$$E = 0.2 + 0.7 + 1.5 = 2.4.$$

Thus,

$$E = 2.4 \text{ expected deaths.}$$

Step 3: Compute SMR

If 3 deaths are observed:

$$\text{SMR} = \frac{3}{2.4} = 1.25.$$

Conclusion

The hospital experienced 25% more deaths than expected based on the reference population.

Example 2: Cancer incidence in a small town (Rates per 100)

Statement

A town of 5000 people has the following age distribution. Reference cancer incidence rates are expressed per 100 people. We want to compute the expected number of cases and the Standardized Mortality Ratio (SMR).

Data

- Total population: 5000

Age Group	N_i	T_i^{ref} (per 100)
0–39	3000	0.1
40–59	1500	0.5
60+	500	2

Step 1: Convert Rates

$$\lambda_{0-39} = 0.001, \quad \lambda_{40-59} = 0.005, \quad \lambda_{60+} = 0.02.$$

Step 2: Compute Expected Cases

$$E = 3000 \times 0.001 + 1500 \times 0.005 + 500 \times 0.02$$

$$E = 3 + 7.5 + 10 = 20.5.$$

Thus,

$$E = 20.5 \text{ expected cases.}$$

Step 3: Compute SMR

If 22 cases are observed:

$$\text{SMR} = \frac{22}{20.5} \approx 1.073.$$

Conclusion

The town experienced approximately 7.3% more cases than expected under the reference rates.

3.5.3 Advantages and Limitations

Advantages

- **Suitable for small populations:** Indirect standardization is ideal when stratum-specific rates in the study population are unstable or based on few events.
- **Does not require reliable study-specific rates:** Only the reference population's rates are needed, making the method robust in situations with sparse data.
- **Produces an interpretable summary measure (SMR):** The Standardized Mortality Ratio provides a clear ratio comparing observed to expected events.
- **Useful for comparing a single study population to a standard:** Commonly applied in occupational epidemiology, hospital performance evaluation, and rare disease studies.
- **Straightforward computation:** Expected events are obtained through simple multiplication of reference rates by study population sizes.

Limitations

- **SMRs are not directly comparable across multiple populations:** Unlike directly standardized rates, SMRs for different study populations cannot be compared unless they share the same reference population.
- **Does not yield an absolute rate:** The SMR is a ratio, not a standardized mortality rate. Therefore, it does not provide an adjusted rate per 1,000 or 100,000.
- **Sensitive to the choice of reference population:** Different reference rates may produce different SMRs, which can complicate interpretation.
- **Assumes reference rates are appropriate for the study population:** If the reference rates do not reflect the underlying risk pattern of the study population, results may be misleading.
- **Cannot directly adjust multiple populations for comparison purposes:** Indirect standardization is mainly used to evaluate one study population relative to a reference, not to compare several populations among themselves.

3.6 Comparison of Methods

- Direct standardization yields **standardized rates**.
- Indirect standardization yields a **ratio** (SMR).
- Choice depends on available data:

- Direct: requires stable stratum-specific rates.
- Indirect: appropriate for small populations.

Exercise: Direct and Indirect Standardization of Mortality Rates

Problem Statement

Two cities (A and B) have different age distributions and observed mortality data. The standard population is provided below.

Age Group	City A Pop	City A Deaths	City B Pop	City B Deaths	Standard Pop
0–44	50,000	100	60,000	150	55,000
45–64	30,000	300	25,000	350	27,500
65+	20,000	600	15,000	500	17,500

Table 3.1: Population and deaths by age group for Cities A and B and the standard population.

1. Compute the Directly Standardized Mortality Rate (DSR) for both cities.
2. Compute the Indirect Standardization for City B using City A as reference (expected cases and SMR).

Solution

1. **Direct Standardization (DSR):**

   The directly standardized mortality rate is defined as:

$$T_{\text{std}}^{\text{direct}} = \frac{\sum_i w_i T_i}{\sum_i w_i}, \quad T_i = \frac{\text{Deaths}_i}{\text{Population}_i}$$

   Step 1a: Age-specific mortality rates (per 1000)

   For City A:

$$T_{0-44,A} = \frac{100}{50{,}000} \times 1000 = 2$$

$$T_{45-64,A} = \frac{300}{30{,}000} \times 1000 = 10$$

$$T_{65+,A} = \frac{600}{20{,}000} \times 1000 = 30$$

   For City B:

$$T_{0-44,B} = \frac{150}{60{,}000} \times 1000 = 2.5$$

$$T_{45-64,B} = \frac{350}{25{,}000} \times 1000 = 14$$

$$T_{65+,B} = \frac{500}{15{,}000} \times 1000 \approx 33.33$$

Step 1b: Apply rates to the standard population

$$DSR_A = \frac{55{,}000 \cdot 2 + 27{,}500 \cdot 10 + 17{,}500 \cdot 30}{100{,}000} = \frac{910{,}000}{100{,}000} = 9.1$$

$$DSR_B = \frac{55{,}000 \cdot 2.5 + 27{,}500 \cdot 14 + 17{,}500 \cdot 33.33}{100{,}000} \approx 11.06$$

Interpretation:
After adjusting for age differences using the standard population, City B has a higher age-adjusted mortality rate (11.06‰) than City A (9.1‰). Thus, independently of age structure, mortality risk is higher in City B.

2. **Indirect Standardization (SMR):**

The expected number of deaths in City B using City A as reference is:

$$E_B = \sum_i N_i \cdot T_i^{\text{ref}}$$

where:

- N_i is the population of City B in age group i,
- T_i^{ref} is the age-specific rate in City A.

Expected deaths in City B:

$$E_B = 60{,}000 \cdot 0.002 + 25{,}000 \cdot 0.01 + 15{,}000 \cdot 0.03 = 120 + 250 + 450 = 820$$

Standardized Mortality Ratio (SMR):
Observed deaths in City B:

$$O_B = 150 + 350 + 500 = 1000$$

$$SMR_B = \frac{O_B}{E_B} = \frac{1000}{820} \approx 1.22$$

Interpretation:
City B experienced approximately 22% more deaths than expected if it had the same age-specific mortality rates as City A.

Practical Work 5: Direct and Indirect Standardization of Mortality Rates in R

Listing 3.1: Direct and Indirect Standardization

```
1
2  # 1. Data
3  age_groups <- c("0-44","45-64","65+")
4
5  pop_A <- c(50000, 30000, 20000)
6  deaths_A <- c(100, 300, 600)
7
8  pop_B <- c(60000, 25000, 15000)
9  deaths_B <- c(150, 350, 500)
10
11 std_pop <- c(55000, 27500, 17500)
12
13 # 2. Step 1: Compute Age-specific Rates (per 1000)
14 rate_A <- deaths_A / pop_A * 1000
15 rate_B <- deaths_B / pop_B * 1000
16
17 # 3. Visualization 1: Age-specific Mortality Rates
18 barplot(
19   rbind(rate_A, rate_B),
20   beside = TRUE,
21   names.arg = age_groups,
22   ylim = c(0, max(rate_A, rate_B) + 5),
23   ylab = "Age-specific mortality rate (per 1000)",
24   col = c("orange", "skyblue"),
25   main = "Visualization 1: Age-specific Mortality Rates"
26 )
27 legend("topleft", legend = c("City A", "City B"),
28        fill = c("orange", "skyblue"))
29
30 # 4. Step 2: Direct Standardization
31 DSR_A <- sum(rate_A * std_pop) / sum(std_pop)
32 DSR_B <- sum(rate_B * std_pop) / sum(std_pop)
33
34 cat("Direct Standardized Rate - City A:", DSR_A, "\n")
35 cat("Direct Standardized Rate - City B:", DSR_B, "\n")
36
37 # 5. Visualization 2: Direct Standardized Rates
38 DSR_values <- c(DSR_A, DSR_B)
39 names(DSR_values) <- c("City A", "City B")
40
41 barplot(
42   DSR_values,
43   col = c("orange", "skyblue"),
44   ylab = "Directly Standardized Rate (per 1000)",
45   main = "Visualization 2: Directly Standardized Rates"
46 )
```

```r
text(x = 1:2, y = DSR_values, labels = round(DSR_values,2), pos = 3)

# 6. Step 3: Indirect Standardization (City B vs City A)
expected_B <- sum(pop_B * (rate_A / 1000))
observed_B <- sum(deaths_B)
SMR_B <- observed_B / expected_B
# ISR_B = Indirect Standardized Rate, derived from SMR for comparison with DSR
# Note: SMR is a ratio (observed/expected) and NOT a directly standardized rate
ISR_B <- SMR_B * DSR_A    # Indirect Standardized Rate

cat("Expected deaths in City B:", expected_B, "\n")
cat("SMR for City B (Observed / Expected):", SMR_B, "\n")
cat("Indirect Standardized Rate (derived from SMR, for comparison with DSR):", ISR_B, "\n")

# 7. Visualization 3: Indirect Standardization
barplot(
  ISR_B,
  names.arg = "City B",
  col = "lightgreen",
  ylab = "Indirect Standardized Rate (per 1000)",
  main = "Visualization 3: Indirect Standardization"
)
text(x = 1, y = ISR_B, labels = round(ISR_B,2), pos = 3)
```

The following chart shows the age-specific mortality rates for City A and City B, highlighting the differences across age groups:

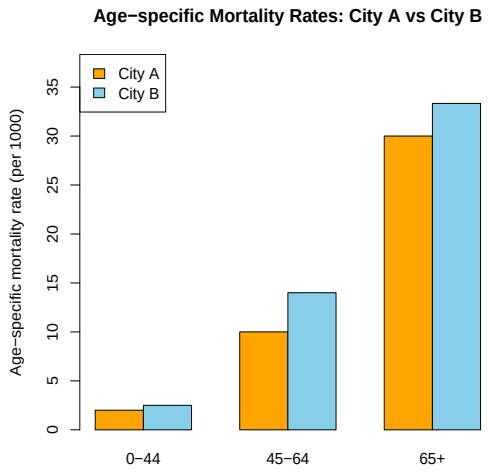


Figure 3.1: Age-specific mortality rates (per 1000) for City A and City B, showing differences across age groups.

Interpretation:
This graph highlights age-related differences in mortality and indicates that direct com-

parisons between cities without standardization may be misleading due to differing age structures.

Using the standard population, we now compute the directly standardized rates for both cities to allow a fair comparison independent of age distribution :

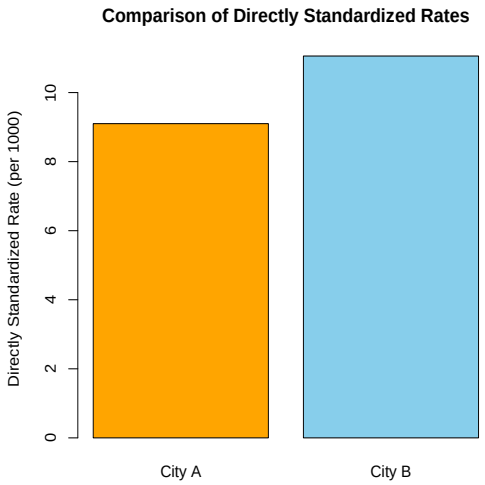


Figure 3.2: Directly Standardized Mortality Rates (DSR) for City A and City B using the standard population.

Interpretation:
Direct standardization shows the "true" difference in mortality between cities, independent of age structure.

Even if City B had higher mortality in younger groups, its overall DSR may be comparable to City A after adjusting for age differences. This is crucial for fair public health comparisons.

Finally, we apply the indirect standardization method, comparing City B to City A as the reference population to obtain the standardized mortality rate.

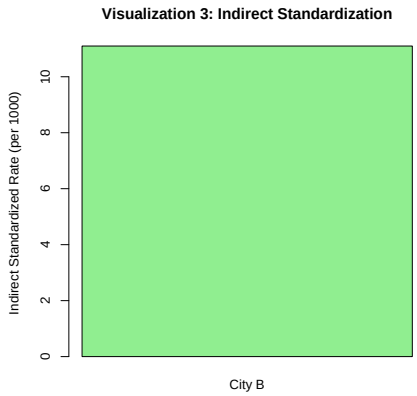


Figure 3.3: Indirect Standardized Mortality Rate (SMR) for City B compared to City A as the reference population.

Interpretation: City B experienced 22% more deaths than expected if it had the same age-specific rates as City A. An SMR greater than 1 indicates an excess risk compared to the reference population, while an SMR less than 1 would indicate lower-than-expected risk.

3.7 Chapter Conclusion

Standardization is an essential methodological tool in epidemiology for producing fair and meaningful comparisons between populations that differ in their demographic structure, particularly age. Crude rates, while easy to compute, can lead to misleading interpretations when populations vary substantially in their composition.

Direct standardization provides a hypothetical rate that applies a population's age-specific rates to a fixed reference distribution. It is most appropriate when the study population has stable and reliable stratum-specific rates, and it allows direct comparison between several populations on a common standard.

Indirect standardization, on the other hand, is particularly valuable when stratum-specific rates in the study population are unstable or unavailable. By applying the reference population's rates to the study population, it yields the Standardized Mortality Ratio (SMR), a powerful measure of relative risk compared to expectations.

Together, these methods complement each other: direct standardization provides comparable adjusted rates, while indirect standardization offers a robust ratio-based evaluation of whether a population experiences more or fewer events than expected. Mastering both approaches is crucial for accurate epidemiological assessment, sound public health decision-making, and the correct interpretation of population health differences.

Transition Note. In chapter 3, we examined methods of standardization to adjust rates and proportions for differences in population structure, enabling fair comparisons across groups. While these techniques provide valuable summary measures, they are limited when we wish to investigate the relationship between individual-level characteristics and binary health outcomes. To move beyond aggregated rates and quantify the effect of covariates on the probability of an event, we now introduce ***logistic regression****, a model that allows estimation of event probabilities, adjustment for multiple predictors simultaneously, and interpretation of effects via odds ratios.*

Chapter 4

Introduction to Logistic Regression

4.1 Motivation and Context

In public health, clinical research, and epidemiology, many variables of interest are inherently binary: a patient is either diseased or not, survives or dies, responds to treatment or does not, smokes or does not smoke. In all these situations, the central scientific question is to estimate and interpret the probability of an event occurring, given relevant predictors such as age, risk factors, or biological measurements. Therefore, statistical models capable of describing how covariates influence a binary outcome are essential for decision-making in medical and epidemiological practice.

Classical linear regression is not suitable for binary outcomes for several reasons. First, the model's predictions are not constrained to the interval $[0, 1]$, leading to impossible probability values (e.g., negative probabilities or values greater than 1). Second, the variance of a binary variable Y given covariates X is $p(1-p)$, which depends on the probability itself; hence the assumption of constant variance (homoscedasticity) required by ordinary least squares (OLS) is violated. Third, because Y takes only the values 0 and 1, the error term $Y - \hat{p}$ cannot follow a normal distribution, which invalidates the distributional assumptions used in linear regression inference. Finally, the relationship between predictors and the event probability is rarely linear in practice: a one-unit increase in a covariate does not have the same impact on the probability when the baseline risk is very small versus when it is already high.

Logistic regression provides a principled and robust solution to these problems by modeling the probability of the event using a nonlinear transformation. Instead of modeling p directly, the model operates on the *log-odds* or *logit* transformation,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

which maps probabilities in $(0, 1)$ to the entire real line. A simple linear model is then assumed for the logit:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Applying the inverse-logit transformation yields the logistic function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}},$$

which guarantees that predicted probabilities always lie strictly between 0 and 1. This formulation has a crucial interpretative advantage: each coefficient β_j corresponds to a

multiplicative effect on the odds of the event. Specifically, e^{β_j} is the *odds ratio* associated with a one-unit increase in X_j, holding the other predictors constant. For example, if $e^{\beta_1} = 3$ for a smoking indicator variable, this means that smokers have three times the odds of developing the disease compared to non-smokers, after adjusting for other factors.

The logistic regression model is not only conceptually appropriate but also statistically well-founded. Because each observation Y_i follows a Bernoulli distribution with parameter p_i, one can construct a likelihood function and estimate the coefficients via maximum likelihood estimation (MLE). This ensures consistent estimators, valid hypothesis tests, confidence intervals, and a flexible framework that naturally integrates into the broader class of generalized linear models (GLMs).

Overall, logistic regression resolves the fundamental limitations of linear regression for binary outcomes, produces meaningful and interpretable measures such as odds ratios, and provides a statistically coherent framework for analyzing binary health outcomes in real-world epidemiological research.

4.2 The Logistic Regression Model: Theoretical Foundations

Logistic regression is designed for situations where the response variable is binary, meaning it can take only two possible outcomes, typically coded as $Y \in \{0, 1\}$.

4.2.1 Binary Outcomes and the Bernoulli Model

Binary outcomes arise naturally in public health and clinical research whenever an event either occurs or does not occur. Examples include whether a patient develops a disease, whether a diagnostic test is positive, or whether a treatment leads to recovery. In all such cases, the response variable Y takes only two possible values:

$$Y = \begin{cases} 1 & \text{if the event of interest occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

This type of variable follows a Bernoulli distribution, which is fully characterized by a single parameter $p = P(Y = 1)$, the probability that the event occurs. The probability mass function is

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

In the context of regression modeling, we assume that the probability of the event may vary from one individual to another depending on observed predictors (covariates) such as age, sex, exposure, clinical measurements, or lifestyle factors. Let $X_{i1}, \ldots, X_{ik}$ denote the predictors for individual i. We write the corresponding probability of the event as

$$p_i = P(Y_i = 1 \mid X_i),$$

which allows each observation to have its own Bernoulli parameter.

Independence Assumption

We assume that the observations (Y_i, X_i), for $i = 1, \ldots, n$, are independent. This structural assumption implies that the joint distribution of the sample is the product of the individual Bernoulli distributions.

For n independent individuals, the joint probability (likelihood) of observing the data $(y_1, \ldots, y_n)$ is then given by

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}.$$

The Bernoulli formulation also highlights two important properties of binary data:

- **Non-constant variance.** The variance of a Bernoulli outcome is

  $$\mathrm{Var}(Y_i \mid X_i) = p_i(1-p_i),$$

  which depends on the mean p_i itself, violating the homoscedasticity assumption required by classical linear regression.

- **Bounded mean.** Because the mean of a Bernoulli variable must lie strictly between 0 and 1, any regression model for p_i must ensure that predicted values always fall within this interval. This is precisely what the logistic (sigmoid) function guarantees.

4.2.2 The Logistic Function and the Logit Link

A central component of logistic regression is the transformation that connects the linear predictor $\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$ to the probability of the event $p_i = P(Y_i = 1 \mid X_i)$. The link function used in logistic regression is the *logit* function:

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

4.2.3 Why the Logit?

The logit is well suited for binary outcomes due to:

- Its range: $(0, 1)$ is mapped to $(-\infty, \infty)$;
- Its symmetry around $p = 0.5$;
- Its interpretability in terms of odds:

  $$\mathrm{odds}(p) = \frac{p}{1-p}, \qquad \mathrm{logit}(p) = \log \mathrm{odds}(p);$$

- Its role as the canonical link of the Bernoulli distribution in GLMs.

4.2.4 The Logistic Regression Equation

Let $X_{i1}, \ldots, X_{ik}$ denote the predictors for individual i. The logistic regression model is

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}.$$

From Logit to Probability

1. Start with the logit model:

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

2. Replace logit with its definition (log-odds):

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

3. Exponentiate both sides to remove the logarithm:

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_k X_{ik}}$$

4. Solve for p_i:

$$\begin{aligned} p_i &= (1-p_i)\cdot e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_k X_{ik}} \\ p_i + p_i \cdot e^{\beta_0+\cdots+\beta_k X_{ik}} &= e^{\beta_0+\cdots+\beta_k X_{ik}} \\ p_i\left(1+e^{\beta_0+\cdots+\beta_k X_{ik}}\right) &= e^{\beta_0+\cdots+\beta_k X_{ik}} \\ p_i &= \frac{e^{\beta_0+\cdots+\beta_k X_{ik}}}{1+e^{\beta_0+\cdots+\beta_k X_{ik}}} \end{aligned}$$

5. Write in standard sigmoid form:

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{i1}+\cdots+\beta_k X_{ik})}}$$

This shows how the linear combination of predictors is transformed into a probability that always lies in the interval $(0, 1)$.

4.2.5 Likelihood and Maximum Likelihood Estimation

Given n independent observations $y_i \in \{0, 1\}$, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

where $p_i = P(Y_i = 1)$. The corresponding log-likelihood is

$$\ell(\beta) = \sum_{i=1}^{n}\Big[y_i \log(p_i) + (1-y_i)\log(1-p_i)\Big].$$

Extension to Logistic Regression: In logistic regression, the probability of success is modeled as

$$p_i(\beta) = \frac{e^{X_i^\top \beta}}{1+e^{X_i^\top \beta}},$$

where X_i is the vector of predictors for observation i.

Step 1: Express $1 - p_i(\beta)$

$$1 - p_i(\beta) = 1 - \frac{e^{X_i^\top \beta}}{1+e^{X_i^\top \beta}} = \frac{1}{1+e^{X_i^\top \beta}}.$$

Step 2: Plug into Bernoulli log-likelihood

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{e^{X_i^\top \beta}}{1+e^{X_i^\top \beta}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{X_i^\top \beta}}\right)\right].$$

Step 3: Simplify logarithms

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n} \Big[y_i(X_i^\top \beta - \log(1+e^{X_i^\top \beta})) + (1-y_i)(-\log(1+e^{X_i^\top \beta}))\Big] \\ &= \sum_{i=1}^{n} \Big[y_i X_i^\top \beta - \log(1+e^{X_i^\top \beta})\Big].\end{aligned}$$

Step 4: Score equations

Starting from the simplified log-likelihood expression,

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i\, X_i^\top \beta - \log\Big(1+e^{X_i^\top \beta}\Big)\right],$$

we differentiate with respect to the parameter vector β.

1. Derivative of the first term. Since

$$\frac{\partial}{\partial \beta}(X_i^\top \beta) = X_i,$$

we obtain

$$\frac{\partial}{\partial \beta}\left(y_i\, X_i^\top \beta\right) = y_i\, X_i.$$

2. Derivative of the second term. Using the chain rule,

$$\frac{\partial}{\partial \beta}\log\Big(1+e^{X_i^\top \beta}\Big) = \frac{e^{X_i^\top \beta}}{1+e^{X_i^\top \beta}}\, X_i,$$

which is equal to

$$p_i(\beta)\, X_i, \qquad \text{where } p_i(\beta) = \frac{e^{X_i^\top \beta}}{1+e^{X_i^\top \beta}}.$$

Thus,

$$\frac{\partial}{\partial \beta}\left[-\log\Big(1+e^{X_i^\top \beta}\Big)\right] = -\,p_i(\beta)\, X_i.$$

3. Combining the derivatives. For each observation i,

$$y_i X_i - p_i(\beta) X_i = X_i\left(y_i - p_i(\beta)\right).$$

4. Final expression of the score function. Summing over all observations yields the score function:

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} X_i\Big(y_i - p_i(\beta)\Big).$$

The maximum likelihood estimator $\widehat{\beta}$ is defined as the solution of the system

$$U(\beta) = 0.$$

Step 5: Computational note

Because the score equations are nonlinear in β, no closed-form solution exists. In practice, numerical optimization methods are required.

The Newton–Raphson algorithm updates the parameter as

$$\beta^{(t+1)} = \beta^{(t)} - H(\beta^{(t)})^{-1} U(\beta^{(t)}),$$

where $H(\beta)$ is the Hessian matrix.

In generalized linear models, this procedure is equivalent to the Iteratively Reweighted Least Squares (IRLS) algorithm. Statistical software such as `R` implements these algorithms efficiently to compute $\widehat{\beta}$ and the associated standard errors.

Hessian Matrix

The Hessian matrix is obtained by differentiating the score function:

$$H(\beta) = \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} X_i X_i^\top \, p_i(1 - p_i).$$

Since $p_i(1 - p_i) > 0$ for all i, the Hessian matrix is negative semi-definite. Therefore, the log-likelihood function is concave, which guarantees that any solution to the score equations corresponds to the unique global maximum.

4.2.6 Interpretation Through the Odds Ratio

In logistic regression, each coefficient β_j corresponds to a multiplicative change in the odds of the event $Y = 1$ for a one-unit increase in the predictor X_j, holding all other variables constant.

The odds for individual i are

$$\text{odds}(p_i) = \frac{p_i}{1 - p_i}.$$

Using the logistic model,

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik},$$

a one-unit increase in X_j changes the log-odds by β_j, which means that the odds are multiplied by

$$\text{OR}_j = e^{\beta_j}.$$

Interpretation of the Odds Ratio

- If $\text{OR}_j > 1$ ($\beta_j > 0$): increasing X_j *increases* the odds of $Y = 1$.
- If $\text{OR}_j < 1$ ($\beta_j < 0$): increasing X_j *decreases* the odds of $Y = 1$.
- If $\text{OR}_j = 1$ ($\beta_j = 0$): increasing X_j *does not change* the odds of $Y = 1$.

Thus, the sign of β_j and the value of the odds ratio e^{β_j} provide a direct and intuitive interpretation of the influence of each predictor.

Example: Odds Ratio Interpretation

Suppose we model the effect of smoking on having a disease using logistic regression, and the estimated coefficient for "smoker" is

$$\beta_{\text{smoker}} = 0.7.$$

Then the odds ratio (OR) is

$$\text{OR} = e^{\beta_{\text{smoker}}} = e^{0.7} \approx 2.01.$$

Interpretation:

- OR > 1 (here 2.01) $\Rightarrow$ smoking *increases* the odds of disease.
- If $\beta_{\text{smoker}} < 0$, e.g., $\beta_{\text{smoker}} = -0.7$, then OR $= e^{-0.7} \approx 0.50 \Rightarrow$ smoking *decreases* the odds.
- If $\beta_{\text{smoker}} = 0$, then OR $= e^0 = 1 \Rightarrow$ smoking has *no effect* on the odds.

Then,

β	OR $= e^{\beta}$	Interpretation
0.7	2.01	Odds increase
-0.7	0.50	Odds decrease
0	1	No effect

4.2.7 Summary of the Theoretical Framework

- Logistic regression models the probability of a binary outcome.
- The logit link transforms probabilities into a linear function.
- Parameters β_j are estimated via maximum likelihood.
- The coefficients provide interpretable odds ratios.
- The log-likelihood is concave, ensuring a unique maximum.

The logistic and logit functions form the mathematical bridge between predictors and event probabilities, ensuring valid probabilities, stable estimates, and interpretable effects.

Exercise: Logistic Regression on a Tiny Dataset

Problem Statement

We consider a study of 6 patients investigating whether regular exercise (X) is associated with disease (Y). To avoid the problem of complete separation, the dataset is constructed so that both groups contain at least one diseased and one healthy individual.

Patient	**Exercise** (X)	**Disease** (Y)
1	0	1
2	0	1
3	0	0
4	1	0
5	1	0
6	1	1

Table 4.1: Dataset for hand-calculation example with Exercise as exposure (revised to avoid separation).

1. Compute the group probabilities of disease.
2. Compute the odds and log-odds (logits).
3. Determine the logistic regression coefficients β_0 and β_1.
4. Compute the odds ratio.
5. Compute the predicted probabilities.
6. Interpret the results.

Solution

1. **Group probabilities:**

   Non-exercisers ($X = 0$)**:**

$$p_0 = \frac{\text{number of diseased}}{\text{total in group}} = \frac{2}{3} \approx 0.667$$

   Exercisers ($X = 1$)**:**

$$p_1 = \frac{1}{3} \approx 0.333$$

2. **Odds and log-odds:**

   Odds

$$\text{Odds}_{X=0} = \frac{p_0}{1 - p_0} = \frac{0.667}{0.333} = 2$$

$$\text{Odds}_{X=1} = \frac{p_1}{1 - p_1} = \frac{0.333}{0.667} = 0.5$$

   Log-odds (logits)

$$\text{logit}(p_0) = \log(2) \approx 0.693$$

$$\text{logit}(p_1) = \log(0.5) = -0.693$$

3. **Logistic regression coefficients:**

   The logistic regression model is:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i$$

   For non-exercisers ($X = 0$):

$$0.693 = \beta_0$$

   For exercisers ($X = 1$):

$$-0.693 = 0.693 + \beta_1$$

$$\beta_1 = -1.386$$

$$\boxed{\beta_0 \approx 0.693, \qquad \beta_1 \approx -1.386}$$

4. **Odds Ratio:**

$$\text{OR} = e^{\beta_1} = e^{-1.386} = 0.25$$

5. **Predicted probabilities:**

   Non-exercisers ($X = 0$):

$$p_0 = \frac{1}{1 + e^{-(\beta_0)}} = \frac{1}{1 + e^{-0.693}} = 0.667$$

   Exercisers ($X = 1$):

$$p_1 = \frac{1}{1 + e^{-(\beta_0 + \beta_1)}} = \frac{1}{1 + e^{0.693}} = 0.333$$

6. **Interpretation:**

- The intercept β_0 represents the log-odds of disease for non-exercisers.
- The coefficient β_1 represents the change in log-odds associated with exercise.
- The odds ratio $e^{\beta_1} = 0.25$ indicates that exercisers have 25% of the odds of disease compared to non-exercisers.
- The predicted probability of disease is 66.7% for non-exercisers and 33.3% for exercisers.

Practical Work 6: Logistic Regression on a Tiny Dataset in R

Listing 4.1: Implementation of Logistic Regression on a Small Illustrative Dataset in R

```
1
2  # Dataset from the worked example
3  X <- c(0, 0, 0, 1, 1, 1)
4  Y <- c(1, 1, 0, 0, 0, 1)
5
6  # Fit logistic regression model
7  model <- glm(Y ~ X, family = binomial(link = "logit"))
8
9  # Extract coefficients
10 beta0 <- coef(model)[1]
11 beta1 <- coef(model)[2]
12
13 # Compute Odds Ratio
14 OR <- exp(beta1)
15
16 # Display results
17 cat("beta0 =", beta0, "\n")
18 cat("beta1 =", beta1, "\n")
19 cat("Odds Ratio =", OR, "\n")
20
21 # Predicted probabilities
22 p_hat <- predict(model, type = "response")
23 cat("Predicted probabilities:", p_hat, "\n")
24
25 # --- Graphical representation ---
26 # Create a grid of X values for the curve
27 x_grid <- seq(0, 1, length = 100)
28
29 # Compute fitted logistic curve
30 logit_curve <- predict(model, newdata = data.frame(X = x_grid),
31                        type = "response")
32
33 # Plot observed data
34 plot(X, Y,
35      xlim = c(-0.1, 1.1), ylim = c(-0.1, 1.1),
36      xlab = "Exercise (X)",
37      ylab = "Disease Probability",
```

```
     main = "Logistic Regression Fit",
     pch = 19)

# Add fitted logistic curve
lines(x_grid, logit_curve, lwd = 2)

# Add fitted probabilities at observed X values
points(X, p_hat, col = "red", pch = 19)
```

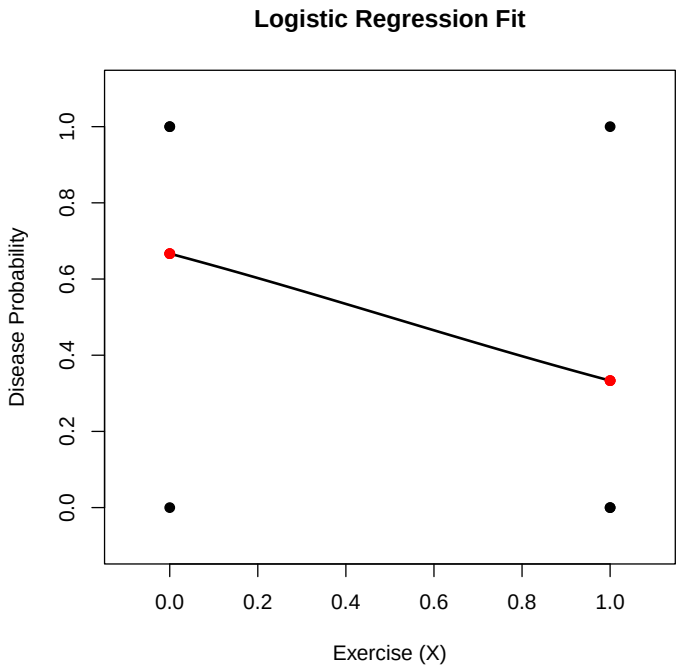


Figure 4.1: Predicted Probability of Disease by Exercise Status from Logistic Regression

Graph Interpretation

The plot shows the relationship between exercise (X) and the probability of disease (Y) as estimated by the logistic regression model.

Y-axis (Disease Probability): This axis represents the *predicted probability of disease* for each patient, ranging between 0 and 1. It does not show the actual disease status (0 or 1), but the probability estimated by the model.

Red points: The red points highlight the predicted probabilities at the observed values of X. They serve to show the model-based prediction clearly for each group.

Fitted logistic curve: The smooth black curve represents the logistic regression function

$$\hat{p}(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}.$$

Because the slope coefficient β_1 is negative, the curve decreases from $X = 0$ to $X = 1$, indicating that exercise is associated with a lower probability of disease.

Overall interpretation:

- The graph illustrates that patients who exercise have a lower predicted probability of disease.

- This visual observation is consistent with the odds ratio:

$$\text{OR} = e^{\beta_1} \approx 0.25,$$

  showing that exercisers have only 25% of the odds of disease compared to non-exercisers.

- The red points on the curve confirm the predicted probabilities for each group, while the black points show the predicted probability for each individual patient.

Exercise: Logistic Regression on Smoking Data

Problem Statement

We consider a study of 8 patients investigating whether smoking (X) is associated with disease (Y).

Patient	**Smoker (X)**	**Disease (Y)**
1	0	0
2	0	0
3	0	1
4	0	0
5	1	1
6	1	1
7	1	1
8	1	0

Table 4.2: Dataset for hand-calculation example: Smoking as exposure.

1. Compute the group probabilities of disease.
2. Compute the odds and log-odds (logits).
3. Determine the logistic regression coefficients β_0 and β_1.
4. Compute the odds ratio.
5. Compute the predicted probabilities.
6. Interpret the results.

Solution

1. **Group probabilities:**

   Non-smokers ($X = 0$):

$$p_0 = \frac{1}{4} = 0.25$$

Smokers ($X = 1$):

$$p_1 = \frac{3}{4} = 0.75$$

2. **Odds and log-odds:**

Odds

$$\text{Odds}_{X=0} = \frac{0.25}{0.75} = \frac{1}{3} \approx 0.333$$

$$\text{Odds}_{X=1} = \frac{0.75}{0.25} = 3$$

Log-odds (logits)

$$\text{logit}(p_0) = \log(0.333) \approx -1.099$$

$$\text{logit}(p_1) = \log(3) \approx 1.099$$

3. **Logistic regression coefficients:**

The logistic regression model is:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i$$

For non-smokers ($X = 0$):

$$-1.099 = \beta_0$$

For smokers ($X = 1$):

$$1.099 = -1.099 + \beta_1$$

$$\beta_1 = 2.198$$

$$\boxed{\beta_0 \approx -1.099, \qquad \beta_1 \approx 2.198}$$

4. **Odds Ratio:**

$$\text{OR} = e^{\beta_1} = e^{2.198} \approx 9$$

5. **Predicted probabilities:**

   Non-smokers ($X = 0$)**:**

$$p_0 = \frac{1}{1 + e^{-(\beta_0)}} = \frac{1}{1 + e^{1.099}} \approx 0.25$$

   Smokers ($X = 1$)**:**

$$p_1 = \frac{1}{1 + e^{-(\beta_0 + \beta_1)}} = \frac{1}{1 + e^{-1.099}} \approx 0.75$$

6. **Interpretation:**

   - The intercept β_0 represents the log-odds of disease for non-smokers.
   - The coefficient β_1 represents the change in log-odds associated with smoking.
   - The odds ratio $e^{\beta_1} \approx 9$ indicates that smokers have 9 times the odds of disease compared to non-smokers.
   - The predicted probability of disease is 25% for non-smokers and 75% for smokers.

Practical Work 7: Logistic Regression on Smoking Data in R

Listing 4.2: Logistic Regression Analysis of Smoking and Disease in R

```
# ------------------------------
# Logistic Regression Example: Smoking vs Disease
# ------------------------------

# 1. Load required libraries
# Install if needed:
# install.packages("ggplot2")
# install.packages("dplyr")
library(ggplot2)
library(dplyr)

# 2. Create dataset
patient <- 1:8
smoker <- c(0,0,0,0,1,1,1,1)
disease <- c(0,0,1,0,1,1,1,0)

data <- data.frame(patient, smoker, disease)

# 3. Fit logistic regression
model <- glm(disease ~ smoker, data = data, family = binomial)
```

```r

# 4. Compute predicted probabilities
data$predicted <- predict(model, type = "response")

# 5. Print data with predicted probabilities
print(data)

# 6. Summarize predicted probabilities by smoking group
data_summary <- data %>%
  group_by(smoker) %>%
  summarise(mean_pred = mean(predicted))

# 7. Plot predicted probabilities as bar chart
ggplot(data_summary, aes(x=factor(smoker), y=mean_pred)) +
  geom_bar(stat="identity", fill="steelblue", alpha=0.6) +
  scale_x_discrete(labels=c("Non-smoker", "Smoker")) +
  ylim(0,1) +
  labs(x="Smoking Status", y="Predicted Probability",
       title="Predicted Probability of Disease by Smoking") +
  theme_minimal()

# 8. Optional: plot individual predicted probabilities
ggplot(data, aes(x=factor(smoker), y=predicted)) +
  geom_point(size=3, color="darkred") +
  scale_x_discrete(labels=c("Non-smoker", "Smoker")) +
  ylim(0,1) +
  labs(x="Smoking Status", y="Predicted Probability",
       title="Individual Predicted Probabilities by Smoking") +
  theme_minimal()
```

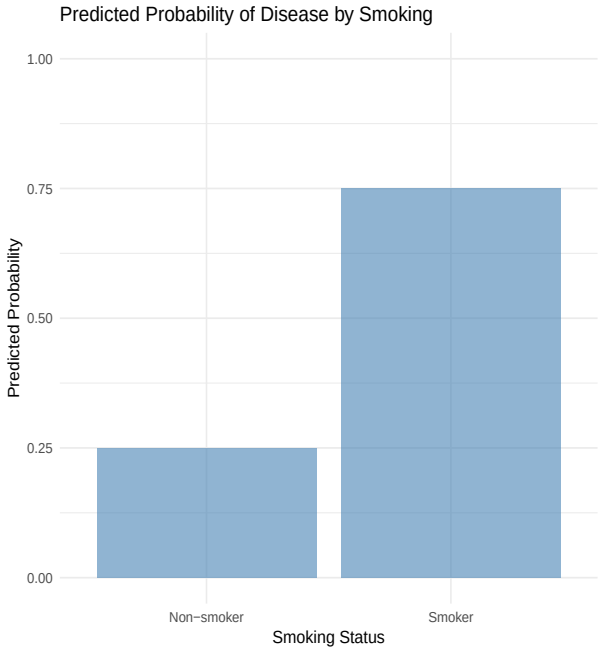


Figure 4.2: Mean predicted probability of disease by smoking status

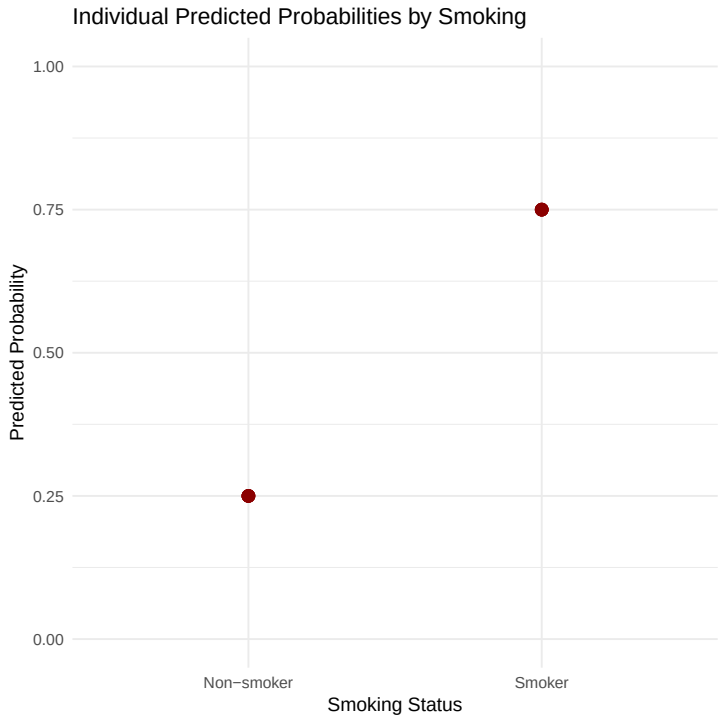


Figure 4.3: Individual predicted probabilities of disease for each patient

Interpretation of the Plots

The first plot shows the mean predicted probability of disease for non-smokers and smokers. Non-smokers have a predicted probability of about 0.25, while smokers have a predicted probability of about 0.75. This indicates that smoking substantially increases the likelihood of having the disease, which is consistent with the odds ratio calculated from the logistic regression model (OR ≈ 9).

The second plot displays the individual predicted probabilities for each patient. Non-smokers cluster around 0.25 and smokers around 0.75, confirming that smokers have higher predicted probabilities of disease. This plot is useful for visualizing individual variation and understanding how the model predicts outcomes for each patient.

Exercise: Maximum Likelihood Estimation in logistic regression

Problem Statement

We study 10 individuals to investigate the effect of `age` and `smoke` on respiratory disease (`disease`).

The dataset is given below.

disease	**age**	**smoke**
1	45	1
0	22	0
1	67	1
0	34	0
0	28	0
1	59	1
0	41	0
1	72	1
0	30	0
1	63	1

Table 4.3: Dataset of 10 individuals: disease status, age, and smoking status.

We denote $x_{i1} = \texttt{age}$ and $x_{i2} = \texttt{smoke}$ for individual i. The logistic regression model is:

$$P(Y_i = 1 \mid X_i) = p_i(\beta) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}$$

The goal is to estimate the parameter vector

$$\beta = (\beta_0, \beta_1, \beta_2)^\top$$

using maximum likelihood estimation (MLE).

1. Write the likelihood function.
2. Derive the log-likelihood.
3. Express the log-likelihood in logistic form.
4. Derive the score equations.
5. Explain why numerical methods are required.

Solution

1. **Likelihood Function:**

   For n independent observations $y_i \in \{0, 1\}$, the Bernoulli likelihood is

$$L(\beta) = \prod_{i=1}^{n} p_i(\beta)^{y_i}(1 - p_i(\beta))^{1-y_i}$$

2. **Log-Likelihood:**

   Taking logarithms, we obtain

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i(\beta)) + (1 - y_i) \log(1 - p_i(\beta))]$$

3. **Logistic Form:**

   Substituting

$$p_i(\beta) = \frac{e^{X_i^\top \beta}}{1 + e^{X_i^\top \beta}}$$

   gives

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i X_i^\top \beta - \log(1 + e^{X_i^\top \beta}) \right]$$

4. **Score Equations:**

   Differentiating with respect to β yields the score function

$$U(\beta) = \sum_{i=1}^{n} X_i (y_i - p_i(\beta))$$

   The maximum likelihood estimator $\hat{\beta}$ satisfies

$$U(\beta) = 0$$

5. **Computational Note:**

   Because the score equations are nonlinear in β, no closed-form solution exists.

   In practice, we compute $\hat{\beta}$ numerically using iterative optimization algorithms such as:

   - Newton–Raphson,
   - Iteratively Reweighted Least Squares (IRLS).

   Software such as `R` implements these procedures efficiently to obtain the maximum likelihood estimates.

Practical Work 8: Maximum Likelihood Estimation in logistic regression in R

Listing 4.3: Logistic Regression Model in R: Maximum Likelihood Estimation with Age and Smoking

```r
# ------------------------------
# Logistic Regression: Age & Smoking
# ------------------------------
library(ggplot2)

# 1. Dataset
data <- data.frame(
  disease = c(1,0,1,0,0,1,0,1,0,1),
  age     = c(45,22,67,34,28,59,41,72,30,63),
  smoke   = c(1,0,1,0,0,1,0,1,0,1)
)

# 2. Fit logistic regression model
model <- glm(disease ~ age + smoke,
             data = data,
             family = binomial)

# 3. Display model summary
summary(model)

# 4. Predicted probabilities for original data
data$predicted <- predict(model, type = "response")

# 5. Create a prediction grid for plotting
age_seq <- seq(min(data$age), max(data$age), length.out = 100)
pred_grid <- expand.grid(
  age = age_seq,
  smoke = c(0,1)
)
pred_grid$predicted <- predict(model, newdata = pred_grid, type =
    "response")

# 6. Plot: Predicted probability by age and smoking status
ggplot() +
  geom_point(data = data, aes(x = age, y = predicted, color =
     factor(smoke)), size = 3) +
  geom_line(data = pred_grid, aes(x = age, y = predicted, color =
      factor(smoke)), linewidth = 1) +
  scale_y_continuous(limits=c(0,1)) +
  scale_color_manual(values=c("0"="blue","1"="red"),
                     labels=c("Non-smoker","Smoker")) +
  labs(
    x = "Age",
    y = "Predicted Probability of Disease",
```

```
    color = "Smoking Status",
    title = "Predicted Probability of Disease by Age and Smoking"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust=0.5))
```

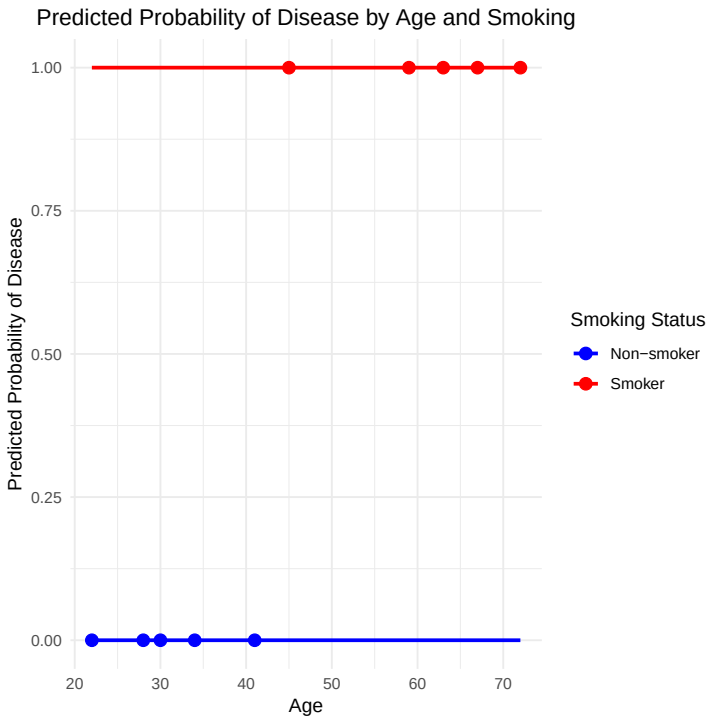


Figure 4.4: Predicted probability of disease by smoking status and age.

Interpretation of the Graph

The plot shows the predicted probability of developing a respiratory disease as a function of `age` for both smokers and non-smokers:

- The probability of disease increases with age for all individuals.
- Smokers (red line) have consistently higher predicted probabilities than non-smokers (blue line) at any given age.
- The gap between smokers and non-smokers highlights the additional risk associated with smoking.
- The individual data points (dots) align well with the predicted probabilities, indicating that the model captures the observed patterns.
- Overall, both predictors (`age` and `smoke`) contribute positively to the risk of disease.

This visualization confirms the model's interpretation: older age and smoking both increase the likelihood of respiratory disease.

4.3 Chapter Conclusion

In this chapter, we introduced logistic regression as a powerful and essential tool for modeling binary outcomes in public health, clinical research, and epidemiology. We highlighted the limitations of classical linear regression when applied to binary data, such as predicted probabilities outside the range $[0, 1]$, non-constant variance, and the violation of normality assumptions.

The chapter emphasized the importance of the Bernoulli distribution as the natural model for binary responses and demonstrated how the logistic regression model links predictors to the probability of an event via the logit function. We explored the interpretation of coefficients in terms of odds ratios, providing a clear and clinically meaningful understanding of the effects of covariates.

Through detailed theoretical examples, we showed how to manually calculate coefficients, interpret the model, and compute predicted probabilities. The practical implementation in R illustrated how these calculations are translated into a real analysis workflow, allowing students to connect theory with computational practice.

Overall, logistic regression provides a statistically coherent framework that ensures valid probability predictions, interpretable results, and meaningful inference. By combining both mathematical understanding and practical implementation, students are equipped to analyze and interpret binary outcomes effectively in epidemiological research.

Transition Note. Chapter 4 focused on logistic regression, linking theory to practical inference with binary outcomes. Building on this foundation, Chapter 5 applies these methods to real epidemiological and medico-economic case studies, emphasizing model implementation, parameter estimation, and decision-making.

Chapter 5

Practical Case Studies: Epidemiological and Medico-Economic Analyses

5.1 General Framework for Applied Case Studies

This chapter serves as a synthesis of the statistical methods introduced in the previous chapters. Rather than presenting new theoretical tools, it focuses on statistical inference in applied settings.

All analyses presented here rely on estimation, hypothesis testing, confidence intervals, and model assessment, applied to epidemiological and medico-economic case studies.

In this chapter, previously introduced statistical methods are applied to complete datasets. Each case study follows a common analytical structure:

1. Specification of observed variables,
2. Formulation of statistical models,
3. Parameter estimation,
4. Numerical interpretation of results.

The focus is exclusively on quantitative analysis.

5.2 Case Study 01: Multivariable Logistic Regression in Epidemiology

Problem Statement

In applied epidemiology, a common challenge is to quantify the relationship between a binary health outcome (e.g., disease presence or absence) and multiple explanatory variables, such as lifestyle factors, demographic characteristics, or clinical measurements. In this study, we consider a dataset of $n = 1{,}500$ individuals with observed variables:

$$Y_i \text{ (disease indicator)}, \quad X_{i1} \text{ (smoking status)}, \quad X_{i2} \text{ (age)}, \quad X_{i3} \text{ (BMI)}.$$

The main research question is: *How do smoking, age, and BMI jointly influence the probability of disease occurrence?*

Objectives

The objectives of this case study are:

1. To fit a multivariable logistic regression model that estimates the probability of disease given multiple covariates.
2. To perform statistical inference on regression coefficients, including hypothesis testing and confidence intervals.
3. To interpret the adjusted effects in terms of odds ratios, providing a clear understanding of the independent contribution of each risk factor.
4. To compute predicted probabilities for specific profiles of individuals, illustrating the practical use of the model.

Connecting Theory with Application

This case study directly applies the theoretical framework developed in the previous sections:

- The **logistic regression equation** and the **logit-to-probability transformation** are used to model the relationship between the covariates and the binary outcome.
- The **maximum likelihood estimation** procedure provides the parameter estimates ($\widehat{\beta}$) that maximize the Bernoulli log-likelihood.
- **Wald tests, standard errors, and confidence intervals** allow us to perform hypothesis testing and quantify uncertainty, as detailed in the theoretical derivation.
- **Odds ratios and predicted probabilities** connect the theoretical interpretation to practical epidemiological insights.

Thus, this section serves as a bridge between abstract theory and applied data analysis, demonstrating how mathematical derivations are realized in a real-world epidemiological study.

Link Between Theoretical Inference and Numerical Implementation in `R`

The inferential results obtained above rely on the numerical maximization of the log-likelihood function and on asymptotic properties of the maximum likelihood estimator. We now establish an explicit bridge between the theoretical framework and its implementation in `R`.

Data Structure and Model Specification in R

Let the observed data be stored in a data frame `data` with variables:

$$Y_i,\ X_{i1},\ X_{i2},\ X_{i3},$$

corresponding respectively to disease status, smoking, age, and BMI.

The logistic regression model

$$\log\left(\frac{p_i(\beta)}{1-p_i(\beta)}\right) = X_i^\top \beta$$

is specified in R using the `glm` function with a binomial family.

```
# Logistic regression model
fit <- glm(
  Y ~ X1 + X2 + X3,
  family = binomial(link = "logit"),
  data = data
)
```

This command instructs R to maximize the Bernoulli log-likelihood

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i X_i^\top \beta - \log(1 + e^{X_i^\top \beta}) \right],$$

using the Iteratively Reweighted Least Squares (IRLS) algorithm.

Numerical Estimation of the Parameter Vector

The estimated regression coefficients correspond to the maximum likelihood estimator $\widehat{\beta}$.

```
# Estimated coefficients
coef(fit)
```

Mathematically, this output provides:

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^\top.$$

Estimated Standard Errors and Asymptotic Variance

The asymptotic covariance matrix of $\widehat{\beta}$ is estimated by the inverse observed Fisher information matrix:

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \mathcal{I}_n(\widehat{\beta})^{-1}.$$

In R, the standard errors are obtained as:

```
# Standard errors of the estimates
summary(fit)$coefficients[, "Std. Error"]
```

These values correspond to $\widehat{\mathrm{SE}}(\widehat{\beta}_j)$ used in the Wald statistics.

Wald Tests and Z-Statistics

Recall that the Wald statistic is defined theoretically as:

$$Z_j = \frac{\widehat{\beta}_j}{\widehat{\mathrm{SE}}(\widehat{\beta}_j)}.$$

In R, the Z-statistics and associated p-values are computed automatically:

```
# Wald test results
summary(fit)
```

The reported `z value` corresponds exactly to the theoretical quantity Z_j, and the `Pr(>|z|)` column provides the asymptotic p-value

$$p_j = 2\left(1 - \Phi(|Z_j|)\right).$$

Odds Ratios and Confidence Intervals

From the theoretical framework, the adjusted odds ratio for covariate X_j is:

$$\widehat{OR}_j = e^{\widehat{\beta}_j}.$$

In R, odds ratios and their confidence intervals are computed as follows:

```
# Odds ratios
exp(coef(fit))

# 95% confidence intervals for odds ratios
exp(confint.default(fit))
```

This corresponds exactly to exponentiating the confidence intervals

$$\widehat{\beta}_j \pm 1.96\,\widehat{\mathrm{SE}}(\widehat{\beta}_j).$$

Predicted Probabilities

Given a new covariate vector

$$X_{\text{new}} = (1, x_1, x_2, x_3)^\top,$$

the predicted probability is:

$$\widehat{p} = \frac{1}{1 + e^{-X_{\text{new}}^\top \widehat{\beta}}}.$$

In R, this is computed using the `predict` function:

```
# New individual
newdata <- data.frame(
  X1 = 1,
  X2 = 50,
  X3 = 27
)

# Predicted probability
predict(fit, newdata, type = "response")
```

The argument `type = "response"` applies the logistic transformation to the linear predictor, yielding $\widehat{p}$.

Synthesis: Theory–Computation Correspondence

Theoretical Quantity	`R` Output
$\ell(\beta)$	IRLS optimization in `glm()`
$\widehat{\beta}$	`coef(fit)`
$\widehat{\mathrm{SE}}(\widehat{\beta})$	`summary(fit)`
Z_j	`z value`
$\widehat{OR}_j$	`exp(coef(fit))`
$\widehat{p}_i$	`predict(..., type="response")`

Inferential Perspective

This numerical implementation confirms that logistic regression inference is entirely driven by maximum likelihood theory. The `R` output provides a direct computational realization of the asymptotic results established in the theoretical framework, thus completing the bridge between mathematical inference and applied data analysis.

5.3 Case Study 02: Effect Modification and Stratified Interpretation

Problem Statement

In applied epidemiology, an important question is whether the effect of a risk factor changes across levels of another variable, a phenomenon known as *effect modification*. In this study, we explore whether the effect of smoking on disease risk depends on age.

The observed data include:

$$Y_i \text{ (disease indicator)}, \quad X_{i1} \text{ (smoking status)}, \quad X_{i2} \text{ (age in years)}.$$

Note: In the `R` code below, the columns are named `Y, X1, X2`, corresponding to Y_i, X_{i1}, X_{i2} respectively.

The main research question is: *Does the effect of smoking on disease probability vary with age?*

Objectives

The objectives of this case study are:

1. To fit a logistic regression model including an interaction term between smoking and age.
2. To estimate the model parameters using maximum likelihood estimation.
3. To interpret the interaction effect and compute age-specific odds ratios.
4. To illustrate the connection between theoretical derivations and numerical implementation in `R`.

Theoretical Model and Maximum Likelihood Estimation

We consider the logistic regression model with interaction:

$$\pi_i = P(Y_i = 1 \mid X_{i1}, X_{i2}) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 (X_{i1} \times X_{i2}))}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 (X_{i1} \times X_{i2}))}.$$

The Bernoulli log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} \Big[y_i (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 (X_{i1} \times X_{i2})) - \log \Big(1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 (X_{i1} \times X_{i2})) \Big) \Big].$$

The maximum likelihood estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^\top$ satisfies the score equations:

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

Estimated Parameters

The estimated coefficients are:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -4.90 \\ 0.60 \\ 0.028 \\ 0.021 \end{pmatrix}.$$

```r
# Fit interaction model in R
fit_inter <- glm(
  Y ~ X1 * X2,
  family = binomial(link = "logit"),
  data = data
)

# Display estimated coefficients
coef(fit_inter)
```

Age-Specific Odds Ratios

The marginal effect of smoking at age a is:

$$\log(OR(a)) = \hat{\beta}_1 + \hat{\beta}_3 a,$$

so that the odds ratio is:

$$OR(a) = \exp(\hat{\beta}_1 + \hat{\beta}_3 a).$$

```r
# Compute age-specific odds ratio, e.g., at age 40
age <- 40
OR_age <- exp(coef(fit_inter)["X1"] + coef(fit_inter)["X1:X2"] *
   age)
OR_age
```

For example, at $a = 40$:

$$OR(40) = \exp(0.60 + 0.021 \times 40) \approx 4.2.$$

Predicted Probabilities for Specific Profiles

Given a new individual with $X_{i1} = 1$ and $X_{i2} = 50$:

$$\eta_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 50 + \hat{\beta}_3 \cdot (1 \times 50)$$

$$\hat{\pi}_{\text{new}} = \frac{1}{1 + e^{-\eta_{\text{new}}}}.$$

```r
# Predicted probability for new individual
newdata <- data.frame(
  X1 = 1,
  X2 = 50
)

predict(fit_inter, newdata, type = "response")
```

Synthesis: Theory–Computation Correspondence

Theoretical Quantity	R Output
$\ell(\beta)$	IRLS optimization in `glm()`
$\hat{\boldsymbol{\beta}}$	`coef(fit_inter)`
Age-specific $OR(a)$	`exp(coef(...) + coef(...)*age)`
Predicted probability $\hat{\pi}_{\text{new}}$	`predict(..., type="response")`

Interpretive Perspective

This case study demonstrates how interaction terms in logistic regression allow the effect of a covariate (smoking) to vary across levels of another variable (age). Maximum likelihood theory provides the parameter estimates, and R offers a practical computational tool to obtain coefficients, age-specific odds ratios, and predicted probabilities, linking theoretical inference to applied epidemiology.

5.4 Case Study 03: Count Data and Overdispersion

Problem Statement

In health services research, it is common to analyze count data, such as the number of hospital admissions per patient over a fixed period. Let

$$Y_i = \text{number of hospital admissions for patient } i \text{ over one year.}$$

A typical approach is to model Y_i using the Poisson distribution. However, in practice, the variance of Y_i often exceeds the mean, a phenomenon known as overdispersion.

The main research question is: *How do patient characteristics influence the number of hospital admissions, accounting for potential overdispersion?*

Objectives

The objectives of this case study are:

1. To fit a Poisson regression model relating count outcomes to covariates.
2. To assess whether overdispersion is present.
3. To adjust standard errors and inference using quasi-likelihood if overdispersion exists.
4. To implement the model and calculations in R.

Poisson Regression Model: Theory

The Poisson regression model assumes:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \log(\lambda_i) = \mathbf{X}_i^\top \boldsymbol{\beta},$$

where $\lambda_i = \mathbb{E}[Y_i \mid \mathbf{X}_i]$.

The log-likelihood is:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i \log(\lambda_i) - \lambda_i - \log(Y_i!) \right].$$

```r
# Example Poisson regression in R
# data contains Y (counts) and covariates X1, X2, X3
fit_pois <- glm(Y ~ X1 + X2 + X3, family = poisson(link = "log"),
    data = data)
summary(fit_pois)
```

Dispersion Check

The dispersion statistic is calculated as:

$$\hat{\phi} = \frac{\text{Deviance}}{n - p},$$

where n is the number of observations and p is the number of parameters.

```r
# Estimated dispersion
deviance(fit_pois) / df.residual(fit_pois)
```

If $\hat{\phi} > 1$, overdispersion is present (as in this case, $\hat{\phi} = 2.4$).

Adjusted Inference Using Quasi-Likelihood

When overdispersion occurs, standard errors are corrected using a quasi-Poisson model:

$$\text{Var}(\hat{\beta}_j)_{\text{adj}} = \hat{\phi} \times \text{Var}(\hat{\beta}_j)_{\text{Poisson}}.$$

```r
# Fit quasi-Poisson model to adjust for overdispersion
fit_quasi <- glm(Y ~ X1 + X2 + X3, family = quasipoisson(link = "
   log"), data = data)
summary(fit_quasi)
```

Interpretive Perspective

- The Poisson regression provides the maximum likelihood estimates of $\boldsymbol{\beta}$.
- The dispersion statistic quantifies whether the Poisson assumption is reasonable.
- If overdispersion is detected, the quasi-Poisson model adjusts the standard errors, ensuring valid Wald tests and confidence intervals.
- R implementation allows direct computation of coefficients, standard errors, deviance, and adjusted inference, bridging theoretical derivations with applied analysis.

5.5 Case Study 04: Medico-Economic Cost-Effectiveness Analysis

Problem Statement

In health economics, a key task is to assess the cost-effectiveness of competing treatments. Given two treatments (A and B), we observe the average cost and average effectiveness (e.g., quality-adjusted life years, QALYs):

Treatment	$\bar{C}$	$\bar{E}$
A	3,800	2.95
B	4,900	3.45

The main research question is: *What is the incremental cost-effectiveness of treatment B relative to A, and how uncertain is this estimate?*

Objectives

The objectives of this case study are:

1. To calculate the incremental cost (ΔC) and incremental effectiveness (ΔE) between treatments.
2. To compute the incremental cost-effectiveness ratio (ICER):
$$ICER = \frac{\Delta C}{\Delta E}.$$
3. To quantify statistical uncertainty of the ICER using the **delta method**.
4. To implement all calculations in `R`.

Incremental Analysis: Theory

The incremental costs and effects are:

$$\Delta C = \bar{C}_B - \bar{C}_A, \quad \Delta E = \bar{E}_B - \bar{E}_A.$$

The ICER is defined as:

$$ICER = \frac{\Delta C}{\Delta E}.$$

```r
# Observed outcomes
cost <- c(3800, 4900)
effect <- c(2.95, 3.45)

# Incremental differences
delta_C <- cost[2] - cost[1]
delta_E <- effect[2] - effect[1]

# ICER
ICER <- delta_C / delta_E
ICER
```

Statistical Uncertainty: Delta Method

Let $(\hat{C}, \hat{E})$ be the vector of sample means for cost and effectiveness, and Σ their covariance matrix.

Define

$$g(C, E) = \frac{C}{E}.$$

The variance of the ICER is approximated via the delta method:

$$\text{Var}(ICER) \approx \nabla g^\top \Sigma \nabla g, \quad \nabla g = \begin{pmatrix} \frac{\partial g}{\partial C} \\ \frac{\partial g}{\partial E} \end{pmatrix} = \begin{pmatrix} 1/E \\ -C/E^2 \end{pmatrix}.$$

```r
# Example covariance matrix (replace with actual sample
    covariance)
Sigma <- matrix(c(100^2, 50, 50, 0.1^2), nrow=2)

# Gradient vector at observed values
grad <- c(1/delta_E, -delta_C/(delta_E^2))

# Variance of ICER
var_ICER <- t(grad) %*% Sigma %*% grad
sd_ICER <- sqrt(var_ICER)
sd_ICER
```

Interpretive Perspective

- The ICER quantifies the additional cost per additional unit of effectiveness for treatment B compared to A.
- The delta method provides a **first-order approximation** of the standard error of the ICER, allowing construction of confidence intervals.
- R implementation computes the ICER, its variance, and standard error, bridging the theoretical derivation with applied analysis.

5.6 Chapter Conclusion

This chapter illustrated how statistical theory is translated into practical data analysis in epidemiology and health economics. Through multiple case studies, we emphasized the following key points:

- Bridging theory and practice: Maximum likelihood estimation, Wald tests, and the delta method provide a theoretical foundation, which can be directly implemented in `R` for applied analyses.
- Model interpretation: Logistic regression models allow quantification of risk factors, including interactions, while Poisson and quasi-Poisson models handle count data and overdispersion. Understanding the meaning of parameters, odds ratios, and predicted probabilities is essential for applied inference.
- Quantifying uncertainty: Confidence intervals, standard errors, and the delta method enable assessment of the reliability of estimates, ensuring that applied conclusions are statistically sound.
- Applied decision-making: In health economics, incremental cost-effectiveness ratios (ICERs) and their uncertainty guide resource allocation, demonstrating the practical relevance of statistical computations.
- Reproducibility and computation: Implementing analyses in `R` bridges the gap between abstract mathematical derivations and real-world datasets, providing reproducible and transparent workflows.

Overall, this chapter reinforced the idea that statistical methods are not merely theoretical constructs but essential tools for understanding health outcomes, informing clinical and policy decisions, and linking quantitative reasoning to actionable insights. By following a structured approach : Defining variables, specifying models, estimating parameters, and interpreting results. Practitioners can confidently conduct applied analyses while maintaining a strong connection to the underlying statistical principles.

Bibliography

[1] Bonita, R., Beaglehole, R., & Kjellström, T. (2006). *Basic Epidemiology* (2nd ed.). World Health Organization.

[2] Clayton, D., & Hills, M. (2013). *Statistical Models in Epidemiology*. Oxford University Press.

[3] Diez, D. M., Barr, C. D., & Çetinkaya-Rundel, M. (2015). *OpenIntro Statistics* (3rd ed.).

[4] Gordis, L. (2013). *Epidemiology* (5th ed.). Elsevier Saunders.

[5] Grolemund, G., & Wickham, H. (2016). *R for Data Science*. O'Reilly Media. Disponible en ligne : https://r4ds.had.co.nz

[6] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

[7] Kirkwood, B. R., & Sterne, J. A. C. (2003). *Essential Medical Statistics* (2nd ed.). Blackwell Science.

[8] Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Wiley.

[9] Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, 9(1), 1–19.

[10] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

[11] Rothman, K. J. (2012). *Epidemiology: An Introduction* (2nd ed.). Oxford University Press.

[12] Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer.

[13] R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Disponible en ligne : https://www.R-project.org/