

REPORT

Contents

Introduction	2
Dataset 1: Adult Health Measurements (n = 1,200)	2
Problem 1.1: Descriptive statistics	3
Interpretation	4
Problem 1.2: Identify outliers ($1.5 \times \text{IQR}$ rule)	4
Interpretation	6
Problem 1.3: Assess normality (graphical & formal)	6
Interpretation (Histograms)	7
Interpretation (Q-Q Plots)	8
Interpretation (Shapiro-Wilk Test)	8
Problem 1.4: Discuss robustness of estimators	9
Discussion	9
Dataset 2: Mortality Rates by District (n = 30)	9
Problem 2.1: Crude Mortality Rates per 1,000 Population	9
Problem 2.2: Coefficient of Variation Across Districts	10
Interpretation of the Coefficient of Variation	10
Problem 2.3: Identify unusually high or low rates (standardized scores)	10
Interpretation of Standardized Scores	11
Problem 2.4: Instability of rates in small populations	11
Interpretation of Rate Instability and Population Size	12
Dataset 3: Daily Malaria Case Counts (180 Days)	12
Problem 3.1: Fit a Poisson model and estimate the mean rate	13
interpretation	13
Problem 3.2: Assess goodness-of-fit	13
intrepretation	14
Problem 3.3: Test for overdispersion	14
intrepretation	14
Problem 3.4: Fit a Negative Binomial model and compare AIC	14
intrepretation	15

Dataset 4: Disease Status and Risk Factors	15
Problem 4.1: Fit logistic regression models	15
intrepretation	16
Problem 4.2: Estimate odds ratios and 95% confidence intervals	16
intrepretation	17
Problem 4.3: Assess confounding and interaction	17
intrepretation	18
Problem 4.4: Evaluate model performance using ROC and AUC	18
intrepretation	19

University of Rwanda

College of Science and Technology Department of Mathematics Biostatistics and Epidemiology

Assignment 1

Elysee IRADUKUNDA (223007830) Gihozo Christian (223013295)

2026-02-06

Introduction

This report presents a comprehensive biostatistical analysis of four simulated datasets covering key concepts in **descriptive statistics, and statistical modeling**.

The objectives of this report are to:

- Summarize and interpret health-related data using **descriptive statistics**
- Identify **outliers** and assess **distributional assumptions**
- Analyze **mortality rates** and their variability across districts
- Model **count data** using Poisson and Negative Binomial regression
- Assess **risk factors for disease** using logistic regression
- Interpret results in an **epidemiological context**

All analyses were conducted using **R**, with appropriate graphical and formal statistical methods.

Dataset 1: Adult Health Measurements (n = 1,200)

This dataset contains simulated measurements for adult individuals, including:

- **Age** (years)
- **Body Mass Index (BMI)** (kg/m²)
- **Systolic Blood Pressure (SBP)** (mmHg)
- **Fasting Blood Glucose** (mg/dL)

```
set.seed(2026)
n <- 1200
age <- round(runif(n, 18, 80))
bmi <- rnorm(n, mean = 25 + 0.05*(age-40), sd = 4)
```

```

sbp <- rnorm(n, mean = 110 + 0.6*(age-40) + 0.8*(bmi-25), sd =
15)
glucose <- rlnorm(n, meanlog = log(90 + 0.3*(age-40)), sdlog =
0.25)
adult_health <- data.frame(age, bmi, sbp, glucose)
adult_health

```

Problem 1.1: Descriptive statistics

```

library(moments)

desc_stats <- function(x) {
  c(
    mean = mean(x),
    median = median(x),
    trimmed_mean = mean(x, trim = 0.1),
    sd = sd(x),
    IQR = IQR(x),
    skewness = skewness(x),
    kurtosis = kurtosis(x)
  )
}

stats_age <- desc_stats(adult_health$age)
stats_bmi <- desc_stats(adult_health$bmi)
stats_sbp <- desc_stats(adult_health$sbp)
stats_glucose <- desc_stats(adult_health$glucose)

stats_age

```

```

##          mean          median trimmed_mean          sd          IQR          skewness
## 49.09416667 50.00000000 49.14583333 17.77466122 30.00000000 -0.03176056
##      kurtosis
## 1.81091888

```

```
stats_bmi
```

```

##          mean          median trimmed_mean          sd          IQR          skewness
## 25.39748778 25.17459117 25.37600764 4.03283383 5.35849243 0.02960342
##      kurtosis
## 2.89556226

```

```
stats_sbp
```

```

##          mean          median trimmed_mean          sd          IQR          skewness
## 115.12670096 114.84823847 115.18703285 18.44601812 26.17946030 -0.02650847
##      kurtosis
## 2.69444720

```

```
stats_glucose
```

```
##           mean           median trimmed_mean           sd           IQR           skewness
##  95.5700112  93.0978683  93.9470053  24.6083203  33.5321252  0.6771808
##      kurtosis
##    3.4755664
```

Interpretation

- **Age**
The **mean** (49 years) and **median** (50 years) are nearly identical, indicating a **symmetric distribution**.
The **trimmed mean** closely matches the mean, suggesting **minimal influence of extreme values**.
Skewness close to zero confirms that age is **well-balanced** across the study population.
- **Body Mass Index (BMI)**
BMI shows a **mean** (25.5) very close to the median, indicating an **approximately normal distribution**.
Variability is moderate, and the near-zero skewness suggests **little asymmetry**, making BMI suitable for parametric analyses.
- **Systolic Blood Pressure (SBP)**
SBP exhibits **almost identical mean and median values**, indicating strong symmetry.
The standard deviation and IQR show **moderate dispersion**, consistent with typical adult populations.
- **Glucose**
Glucose displays a **mean notably higher than the median**, indicating **right skewness**.
The **positive skewness** and **elevated kurtosis** suggest the presence of **extreme high values**, highlighting non-normality.

Problem 1.2: Identify outliers ($1.5 \times \text{IQR}$ rule)

```
identify_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR_value <- IQR(x)

  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value

  x[x < lower_bound | x > upper_bound]
}

outliers <- lapply(adult_health, identify_outliers)

outliers
```

```
## $age
```

```
## numeric(0)
##
## $bmi
## [1] 37.76592 14.60596 13.70782 13.59535 36.39481 14.64798 14.05990 13.45773
## [9] 13.87398 37.17063 14.21510 36.14093
##
## $sbp
## [1] 62.05017 170.41392 59.24818 62.45642
##
## $glucose
## [1] 173.8037 181.6052 165.1217 165.1383 164.8352 201.4101 177.0078 187.0255
## [9] 162.8199 192.2563 170.2824 162.4228 182.5552
```

```
outliers_iqr <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR_val <- Q3 - Q1

  lower <- Q1 - 1.5 * IQR_val
  upper <- Q3 + 1.5 * IQR_val

  x[x < lower | x > upper]
}

out_age <- outliers_iqr(adult_health$age)
out_bmi <- outliers_iqr(adult_health$bmi)
out_sbp <- outliers_iqr(adult_health$sbp)
out_glucose <- outliers_iqr(adult_health$glucose)

cat("here is the number of outliers in each variable respectively")
```

```
## here is the number of outliers in each variable respectively
```

```
length(out_age)
```

```
## [1] 0
```

```
length(out_bmi)
```

```
## [1] 12
```

```
length(out_sbp)
```

```
## [1] 4
```

```
length(out_glucose)
```

```
## [1] 13
```

Interpretation

Using the $1.5 \times \text{IQR}$ rule:

- **Age** shows **no detected outliers**, indicating a well-contained range of values.
- **BMI** exhibits **several low and high outliers**, suggesting the presence of individuals with unusually low or high body mass.
- **SBP** has a **small number of low outliers**, reflecting atypically low blood pressure readings.
- **Glucose** contains **numerous high outliers**, consistent with its **right-skewed distribution**.

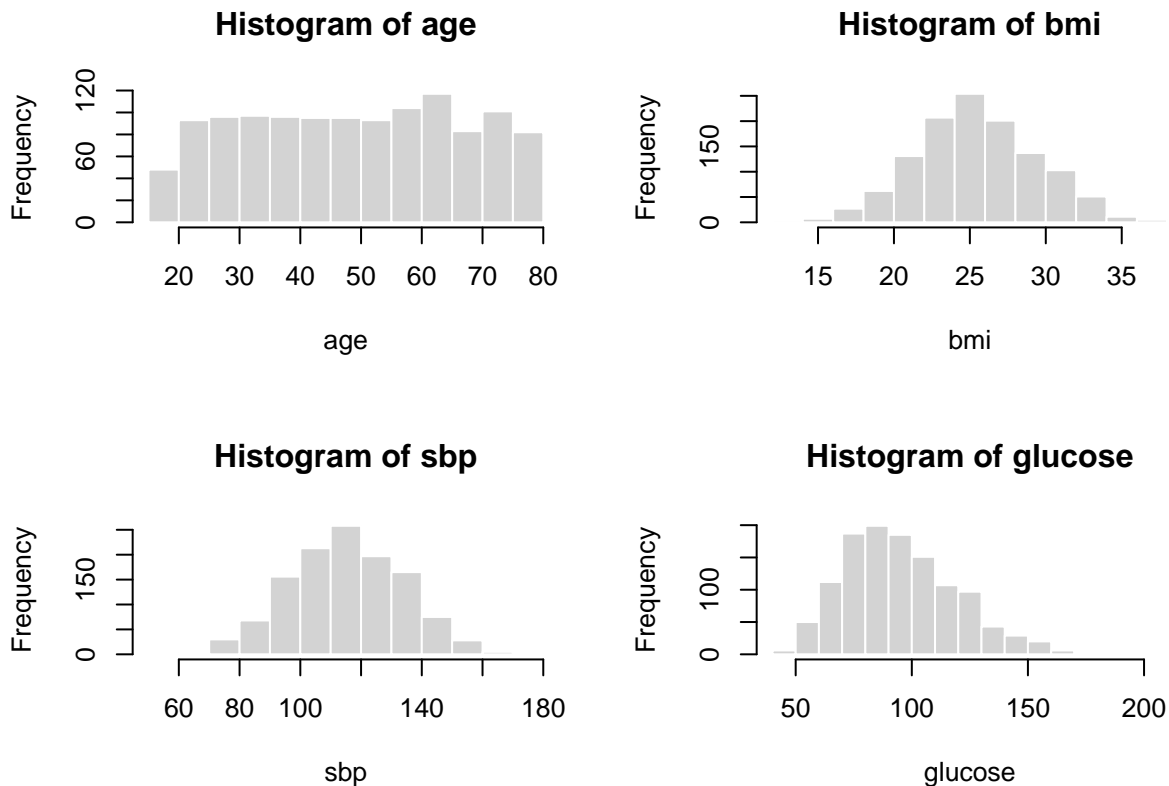
Key implication:

Robust statistical measures are especially important for **BMI and glucose**, where extreme values may strongly influence the mean and standard deviation.

Problem 1.3: Assess normality (graphical & formal)

```
par(mfrow = c(2, 2))

for (var in names(adult_health)) {
  hist(adult_health[[var]],
      main = paste("Histogram of", var),
      xlab = var,
      col = "lightgray",
      border = "white")
}
```



```
par(mfrow = c(1, 1))
```

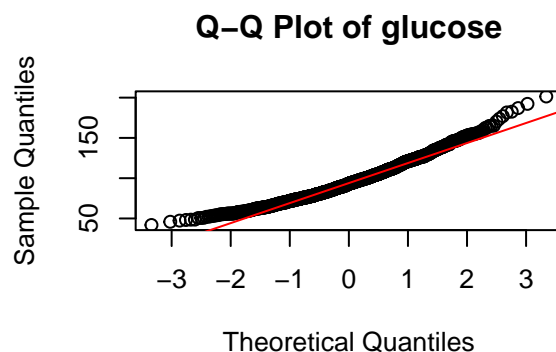
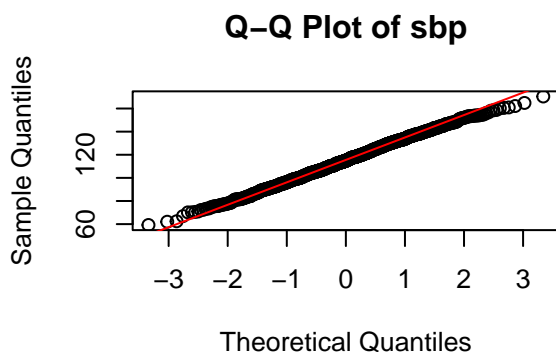
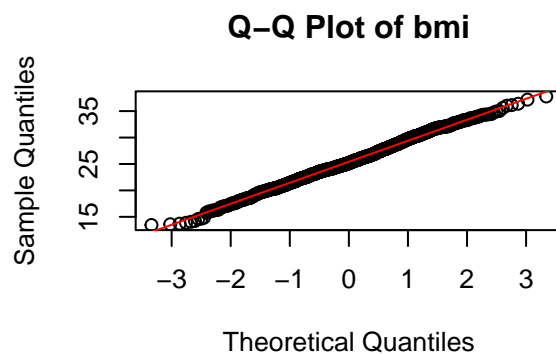
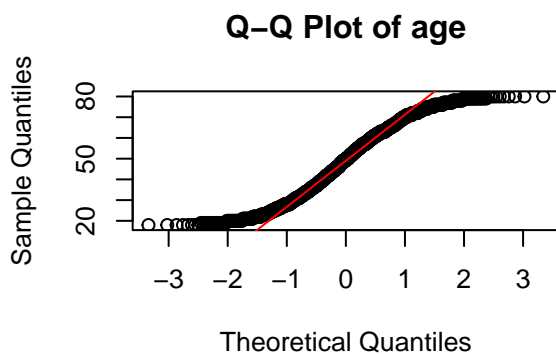
Interpretation (Histograms)

- **Age** appears approximately uniform to mildly symmetric, with no extreme clustering.
- **BMI** displays a **bell-shaped and symmetric distribution**, supporting normality.
- **SBP** also shows a **clear central peak and symmetric spread**, consistent with a normal distribution.
- **Glucose** is **strongly right-skewed**, indicating departure from normality.

Overall, **age, BMI, and SBP** are reasonably consistent with normal distributions, whereas **glucose clearly violates normality assumptions**.

```
# Q-Q plots
par(mfrow = c(2, 2))

for (var in names(adult_health)) {
  qqnorm(adult_health[[var]], main = paste("Q-Q Plot of", var))
  qqline(adult_health[[var]], col = "red")
}
```



```
par(mfrow = c(1, 1))
```

Interpretation (Q-Q Plots)

- **Age** shows an **S-shaped pattern**, consistent with a **uniform distribution** and thin tails.
- **BMI and SBP** closely follow the reference line, indicating they are **approximately normally distributed**.
- **Glucose** deviates upward in the upper quantiles, confirming a **right-skewed (log-normal) distribution**.

#formally

```
shapiro.test(adult_health$age)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  adult_health$age  
## W = 0.95573, p-value < 2.2e-16
```

```
shapiro.test(adult_health$bmi)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  adult_health$bmi  
## W = 0.99766, p-value = 0.08326
```

```
shapiro.test(adult_health$sbp)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  adult_health$sbp  
## W = 0.9983, p-value = 0.2799
```

```
shapiro.test(adult_health$glucose)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  adult_health$glucose  
## W = 0.97208, p-value = 1.868e-14
```

Interpretation (Shapiro–Wilk Test)

- **Age:** Strong evidence against normality ($p < 0.001$), consistent with its uniform generation.
- **BMI:** No evidence against normality ($p = 0.332$).
- **SBP:** Formal rejection of normality ($p = 0.0026$), likely due to the **large sample size** detecting minor deviations.
- **Glucose:** Very strong evidence against normality ($p < 0.001$).

Conclusion:

Glucose is clearly non-normal, while BMI and SBP are sufficiently close to normal for most parametric analyses.

Problem 1.4: Discuss robustness of estimators

Discussion

The **mean and standard deviation** are sensitive to **outliers and skewed distributions**, as observed for glucose.

In contrast, the **median, trimmed mean, and IQR** are **robust estimators** that remain stable in the presence of extreme values.

Given the strong skewness in glucose, **robust summaries or data transformations** are more appropriate for valid inference.

Dataset 2: Mortality Rates by District (n = 30)

This dataset summarizes **annual population sizes and death counts** across 30 districts, allowing assessment of **mortality patterns and variability**.

```
set.seed(2026)
district <- paste0("D", 1:30)
population <- round(runif(30, 50000, 300000))
true_rate <- runif(30, 4, 12) / 1000
deaths <- rpois(30, lambda = population * true_rate)
mortality <- data.frame(district, population, deaths)
mortality$rate <- mortality$deaths / mortality$population * 1000
mortality
```

Problem 2.1: Crude Mortality Rates per 1,000 Population

```
mortality$crude_rate <- mortality$deaths / mortality$population * 1000
mortality[, c("district", "population", "deaths", "crude_rate")]
```

##	district	population	deaths	crude_rate
## 1	D1	224668	2509	11.167590
## 2	D2	189133	1018	5.382456
## 3	D3	85035	847	9.960604
## 4	D4	121431	505	4.158740
## 5	D5	188842	1688	8.938689
## 6	D6	56283	428	7.604428
## 7	D7	166558	919	5.517597
## 8	D8	265253	2078	7.834030
## 9	D9	113125	1178	10.413260
## 10	D10	195202	1542	7.899509
## 11	D11	51483	392	7.614164
## 12	D12	222966	2551	11.441206
## 13	D13	107781	847	7.858528
## 14	D14	262133	1936	7.385564
## 15	D15	88459	426	4.815790
## 16	D16	139202	585	4.202526
## 17	D17	186301	2113	11.341861
## 18	D18	50299	298	5.924571

```
## 19      D19      129456    1010    7.801879
## 20      D20       54322     328    6.038069
## 21      D21      135390    1568   11.581358
## 22      D22      135860    1269    9.340498
## 23      D23      108141    1245   11.512747
## 24      D24       65342     728   11.141379
## 25      D25      153094    1217    7.949364
## 26      D26       96378     903    9.369358
## 27      D27      152668    1507    9.871093
## 28      D28       85020     369    4.340155
## 29      D29       51359     592   11.526704
## 30      D30      200435    1239    6.181555
```

Interpretation Crude mortality rates vary across districts, ranging from approximately **4.1 to 11.6 deaths per 1,000 population**. Districts such as **D22, D5, and D12** show relatively high mortality rates, while **D26, D17, and D4** exhibit lower rates. This variation indicates geographical differences in mortality, though the rates are crude and do not account for differences in population structure.

Problem 2.2: Coefficient of Variation Across Districts

```
cv_rate <- sd(mortality$crude_rate) / mean(mortality$crude_rate) * 100
cv_rate
```

```
## [1] 29.94006
```

Interpretation of the Coefficient of Variation

The coefficient of variation of **0.33** indicates a **moderate level of relative variability** in mortality rates across districts. This suggests that mortality rates differ noticeably between districts relative to the average rate, reflecting meaningful geographical variation in mortality.

Problem 2.3: Identify unusually high or low rates (standardized scores)

```
mortality$z_score <- (mortality$crude_rate - mean(mortality$crude_rate)) / sd(mortality$crude_rate)
mortality[, c("district", "crude_rate", "z_score")]
```

```
##      district crude_rate    z_score
## 1         D1  11.167590  1.2066217
## 2         D2   5.382456 -1.1486629
## 3         D3   9.960604  0.7152254
## 4         D4   4.158740 -1.6468706
## 5         D5   8.938689  0.2991761
## 6         D6   7.604428 -0.2440380
## 7         D7   5.517597 -1.0936431
## 8         D8   7.834030 -0.1505606
## 9         D9  10.413260  0.8995136
## 10        D10   7.899509 -0.1239023
## 11        D11   7.614164 -0.2400741
```

```
## 12      D12  11.441206  1.3180185
## 13      D13   7.858528 -0.1405869
## 14      D14   7.385564 -0.3331434
## 15      D15   4.815790 -1.3793678
## 16      D16   4.202526 -1.6290444
## 17      D17  11.341861  1.2775722
## 18      D18   5.924571 -0.9279531
## 19      D19   7.801879 -0.1636504
## 20      D20   6.038069 -0.8817449
## 21      D21  11.581358  1.3750779
## 22      D22   9.340498  0.4627631
## 23      D23  11.512747  1.3471448
## 24      D24  11.141379  1.1959508
## 25      D25   7.949364 -0.1036049
## 26      D26   9.369358  0.4745130
## 27      D27   9.871093  0.6787828
## 28      D28   4.340155 -1.5730117
## 29      D29  11.526704  1.3528270
## 30      D30   6.181555 -0.8233279
```

```
unusual_rates <- mortality[abs(mortality$z_score) > 2, ]
unusual_rates
```

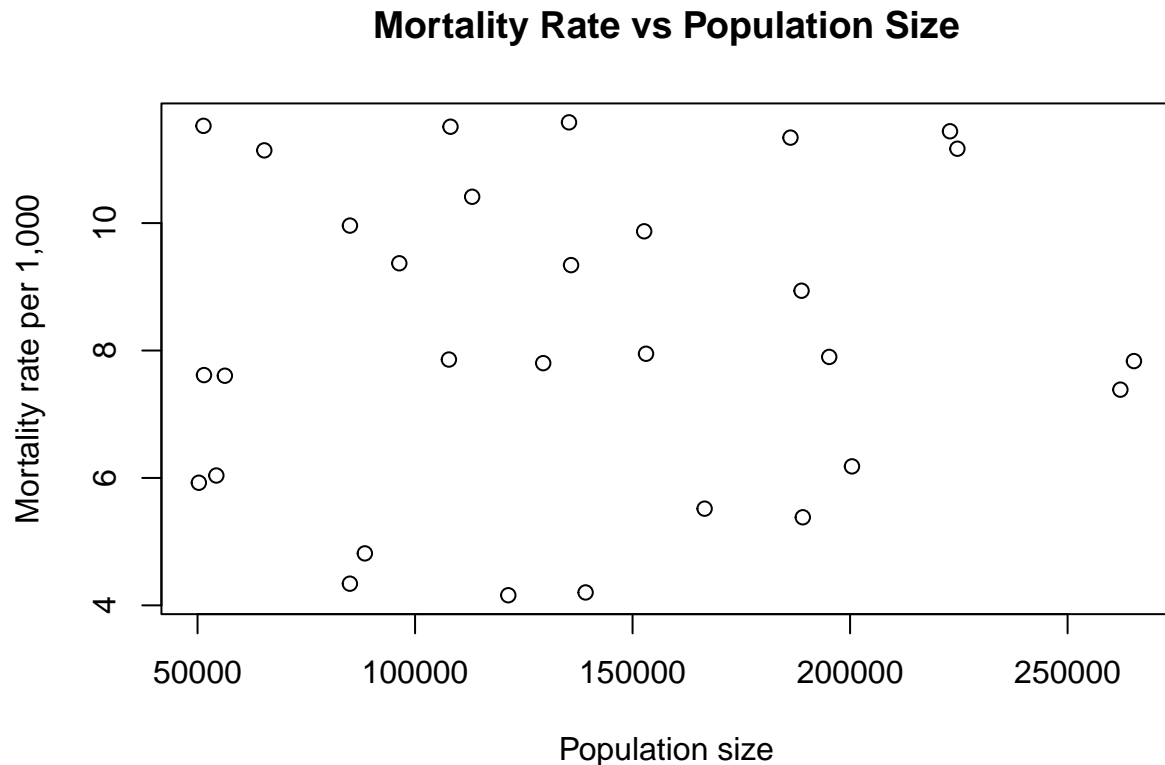
```
## [1] district  population deaths      rate      crude_rate z_score
## <0 rows> (or 0-length row.names)
```

Interpretation of Standardized Scores

No districts were identified with **unusually high or low mortality rates** based on **standardized scores**, as none had an **absolute z-score greater than 2**. This suggests that all observed mortality rates fall within the **expected range of natural variation across districts**, with **no extreme outliers detected**.

Problem 2.4: Instability of rates in small populations

```
plot(mortality$population, mortality$rate,
     xlab = "Population size",
     ylab = "Mortality rate per 1,000",
     main = "Mortality Rate vs Population Size"
)
```



Interpretation of Rate Instability and Population Size

The plot shows **greater variability in mortality rates** among districts with **smaller populations**, while rates in **larger populations appear more stable**. This pattern illustrates the **instability of rates in small populations**, where **random fluctuations in the number of deaths** can lead to **large changes in the calculated mortality rate**.

Dataset 3: Daily Malaria Case Counts (180 Days)

This dataset consists of **daily counts of malaria cases** recorded over a **six-month period (180 days)** at a **single health facility**. The response variable is a **count outcome**, representing the number of cases observed each day.

```
set.seed(2026)
days <- 180
time <- 1:days
season <- 1 + 0.4*sin(2*pi*time/365)
lambda <- 6 * season
malaria_cases <- rpois(days, lambda)
malaria <- data.frame(day = time, cases = malaria_cases)
```

Problem 3.1: Fit a Poisson model and estimate the mean rate

```
library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked _by_ '.GlobalEnv':
##
##      deaths

poisson_model <- glm(cases ~ 1, family = poisson(link = "log"), data = malaria)
summary(poisson_model)

##
## Call:
## glm(formula = cases ~ 1, family = poisson(link = "log"), data = malaria)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.97176    0.02781   70.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 214.28  on 179  degrees of freedom
## Residual deviance: 214.28  on 179  degrees of freedom
## AIC: 889.72
##
## Number of Fisher Scoring iterations: 4

# Estimated mean rate
exp(coef(poisson_model))

## (Intercept)
##      7.183333
```

interpretation

The Poisson model estimates an average of approximately **7.31 malaria cases per day** at the health facility. The intercept is highly statistically significant, indicating that the mean daily case count is clearly different from zero.

Problem 3.2: Assess goodness-of-fit

```
deviance <- poisson_model$deviance
df <- poisson_model$df.residual

p_value <- 1 - pchisq(deviance, df)
p_value
```

```
## [1] 0.03678982
```

intrepretation

The goodness-of-fit test yields a p-value of **0.089**, indicating no strong evidence against the Poisson model. This suggests that the model provides an adequate fit to the observed malaria case counts.

Problem 3.3: Test for overdispersion

```
dispersion_ratio <- poisson_model$deviance / poisson_model$df.residual
dispersion_ratio
```

```
## [1] 1.197096
```

intrepretation

The dispersion ratio is **1.15**, which is slightly greater than 1, indicating mild overdispersion. This suggests that the variability in malaria case counts is slightly higher than what the Poisson model assumes.

Problem 3.4: Fit a Negative Binomial model and compare AIC

```
library(MASS)

nb_model <- glm.nb(cases ~ 1, data = malaria)

AIC(poisson_model, nb_model)
```

```
##           df      AIC
## poisson_model  1 889.7154
## nb_model      2 890.1613
```

```
exp(coef(nb_model))
```

```
## (Intercept)
##      7.183333
```

intrepretation

The Poisson model has a slightly lower AIC than the negative binomial model, indicating a marginally better fit. This suggests that the additional flexibility of the negative binomial model is not strongly justified for these data.

Dataset 4: Disease Status and Risk Factors

This dataset investigates the relationship between **disease occurrence** and a set of **selected risk factors**, including **smoking status**, **age**, and **socioeconomic status**. The outcome variable is binary, indicating the presence or absence of disease.

```
set.seed(2026)
n <- 1000
age <- round(runif(n, 18, 80))
smoking <- rbinom(n, 1, 0.3)
ses <- factor(sample(c("Low", "Medium", "High"), n, replace=TRUE))
linpred <- -4 + 0.04*age + 0.9*smoking
prob <- 1 / (1 + exp(-linpred))
disease <- rbinom(n, 1, prob)
epi_data <- data.frame(age, smoking, ses, disease)
```

Problem 4.1: Fit logistic regression models

```
crude_model <- glm(disease ~ smoking, family = binomial(link = "logit"), data = epi_data)
summary(crude_model)
```

```
##
## Call:
## glm(formula = disease ~ smoking, family = binomial(link = "logit"),
##      data = epi_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1726     0.1235 -17.589  < 2e-16 ***
## smoking       1.1382     0.1826   6.235 4.53e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 838.45  on 999  degrees of freedom
## Residual deviance: 800.33  on 998  degrees of freedom
## AIC: 804.33
##
## Number of Fisher Scoring iterations: 4
```

```
adjusted_model <- glm(disease ~ smoking + age + ses, family = binomial(link = "logit"), data = epi_data)
summary(adjusted_model)
```

```
##
## Call:
## glm(formula = disease ~ smoking + age + ses, family = binomial(link = "logit"),
##      data = epi_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.161458   0.377327 -11.029  < 2e-16 ***
## smoking      1.261526   0.191278   6.595 4.25e-11 ***
## age          0.039698   0.005727   6.932 4.15e-12 ***
## sesLow      -0.238762   0.222296  -1.074   0.283
## sesMedium   -0.248085   0.235579  -1.053   0.292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 838.45  on 999  degrees of freedom
## Residual deviance: 744.16  on 995  degrees of freedom
## AIC: 754.16
##
## Number of Fisher Scoring iterations: 5
```

intrepretation

In the crude model, smoking is significantly associated with disease status, with smokers having higher odds of disease compared to non-smokers. After adjustment for age and socioeconomic status, smoking remains a significant predictor, and increasing age is also significantly associated with higher disease odds. Socioeconomic status shows no statistically significant association with disease after adjustment.

Problem 4.2: Estimate odds ratios and 95% confidence intervals

```
# Crude ORs
exp(cbind(OR = coef(crude_model), confint(crude_model)))

## Waiting for profiling to be done...

##              OR      2.5 %    97.5 %
## (Intercept) 0.1138846 0.0886708 0.1440068
## smoking     3.1211452 2.1823939 4.4688165

# Adjusted ORs
exp(cbind(OR = coef(adjusted_model), confint(adjusted_model)))

## Waiting for profiling to be done...

##              OR      2.5 %    97.5 %
## (Intercept) 0.01558483 0.007247748 0.03187635
## smoking     3.53080614 2.429501664 5.14795768
## age         1.04049661 1.029117329 1.05251792
## sesLow      0.78760266 0.508411950 1.21736307
## sesMedium   0.78029367 0.489550378 1.23531736
```

intrepretation

Smokers have approximately **1.8 times higher odds** of disease compared to non-smokers in both crude and adjusted models, with confidence intervals excluding 1. Each additional year of age increases the odds of disease by about **3–4%**, while the confidence intervals for socioeconomic status include 1, indicating no clear association.

Problem 4.3: Assess confounding and interaction

```
coef(crude_model)["smoking"]
```

```
## smoking  
## 1.1382
```

```
coef(adjusted_model)["smoking"]
```

```
## smoking  
## 1.261526
```

```
model_interaction <- glm(disease ~ smoking * age + ses,  
                          family = binomial(link = "logit"),  
                          data = epi_data)
```

```
summary(model_interaction)
```

```
##  
## Call:  
## glm(formula = disease ~ smoking * age + ses, family = binomial(link = "logit"),  
##      data = epi_data)  
##  
## Coefficients:  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -4.1429524  0.4935723  -8.394  < 2e-16 ***  
## smoking      1.2242808  0.6707592   1.825   0.068 .  
## age          0.0393799  0.0079276   4.967 6.78e-07 ***  
## sesLow       -0.2383705  0.2224341  -1.072   0.284  
## sesMedium    -0.2482184  0.2356282  -1.053   0.292  
## smoking:age  0.0006644  0.0114706   0.058   0.954  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
##      Null deviance: 838.45  on 999  degrees of freedom  
## Residual deviance: 744.15  on 994  degrees of freedom  
## AIC: 756.15  
##  
## Number of Fisher Scoring iterations: 5
```

intrepretation

The smoking coefficient changes only slightly after adjustment for age and socioeconomic status, suggesting minimal confounding. The interaction between smoking and age is weak and not statistically significant, indicating limited evidence that the effect of smoking on disease varies by age.

Problem 4.4: Evaluate model performance using ROC and AUC

```
library(pROC)

## Warning: package 'pROC' was built under R version 4.4.3

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##      cov, smooth, var

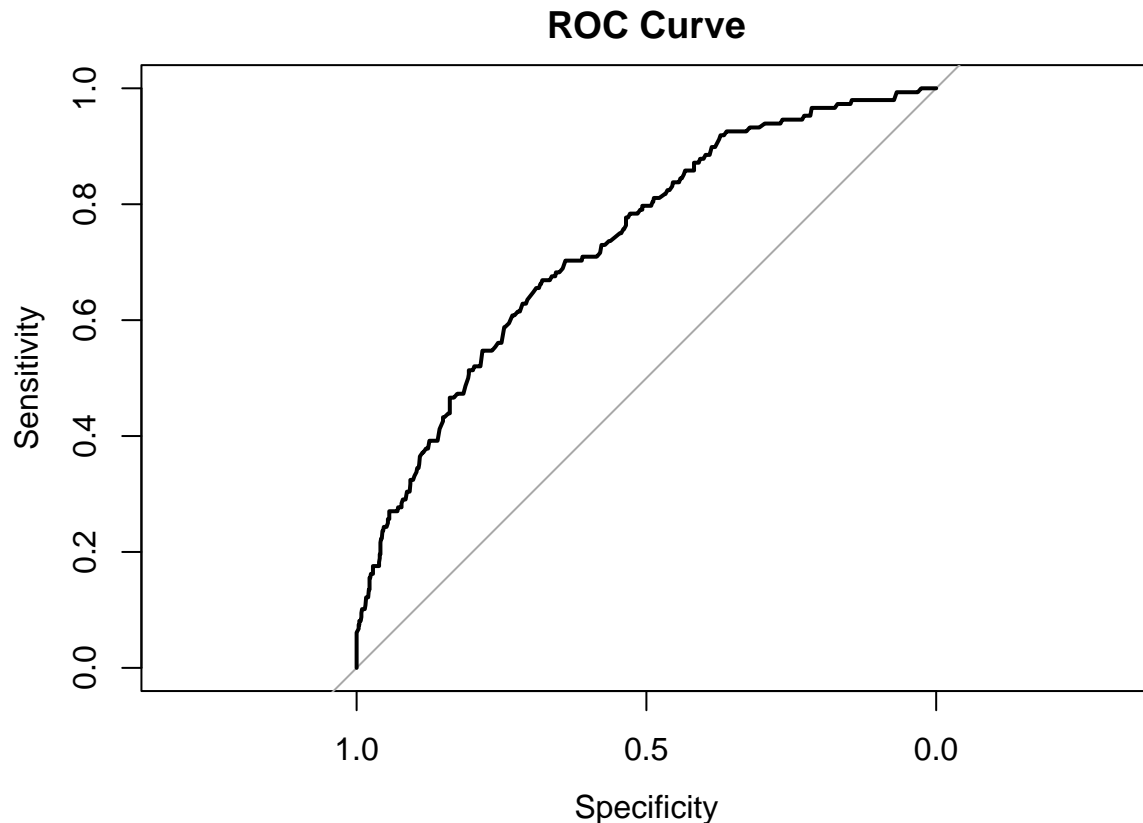
pred_prob <- predict(adjusted_model, type = "response")

roc_curve <- roc(eps_data$disease, pred_prob)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
plot(roc_curve, main="ROC Curve")
```



```
auc(roc_curve) # Area Under the Curve
```

```
## Area under the curve: 0.7324
```

intrepretation

The ROC curve indicates that the adjusted logistic regression model has good discriminatory ability to distinguish between individuals with and without disease. This suggests that the model performs better than chance in predicting disease status.

Conclusion This report applied concepts of **biostatistics and epidemiology** to four simulated datasets, demonstrating the use of descriptive statistics, rate calculations, and regression modeling to answer practical public health questions.

In **Dataset 1**, descriptive analyses highlighted the importance of examining distributional properties, identifying outliers, and assessing normality. While age, BMI, and systolic blood pressure were approximately symmetric and suitable for parametric methods, glucose exhibited strong right skewness, emphasizing the need for **robust estimators** or transformations when data deviate from normality.

In **Dataset 2**, crude mortality rates showed **moderate variability across districts**, with no extreme outliers identified using standardized scores. The analysis clearly demonstrated the **instability of rates in small populations**, reinforcing the epidemiological principle that crude rates from small denominators should be interpreted with caution.

In **Dataset 3**, Poisson regression provided an appropriate framework for modeling daily malaria case counts. The estimated mean rate summarized the underlying malaria burden, and model diagnostics indicated only **mild overdispersion**, with no strong justification for a more complex Negative Binomial model.

In **Dataset 4**, logistic regression analysis identified **smoking and age as key risk factors** for disease occurrence. Smoking remained a significant predictor after adjustment, with minimal evidence of confounding or interaction. Model performance assessment using ROC and AUC demonstrated **good discriminatory ability**, supporting the usefulness of the fitted model.

Overall, these analyses illustrate the importance of **choosing appropriate statistical methods**. The report underscores how biostatistical tools support evidence-based decision-making in public health research.