Assignment part 1: Cleveland heart disease dataset

R version

Introduction

In this assignment, you will work with an individualized dataset derived from the Cleveland heart disease dataset. This dataset contains information on patients with suspected heart disease and includes various demographic, clinical, and diagnostic variables. Your task is to perform a series of analyses to explore the dataset and investigate the relationship between different variables and the presence of heart disease.

Dataset description

The Cleveland heart disease dataset originates from the Cleveland Clinic Foundation and focuses on heart disease diagnosis. It includes data from 303 patients on the following variables:

- `age`: Age in years
- `sex`: Sex (1 = female; 2 = male)
- `cp`: Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- `trestbps`: Resting blood pressure (mm Hg at hospital admission)
- `chol`: Serum cholesterol in mg/dl
- `fbs`: Fasting blood sugar > 120 mg/dl
- `restecg`: Resting electrocardiographic results (1 = normal; 2 = ST-T wave abnormality; 3 = left ventricular hypertrophy)
- `thalach`: Maximum heart rate achieved
- `exang`: Exercise-induced angina (1 = no; 2 = yes)
- `oldpeak`: ST depression induced by exercise relative to rest
- `slope`: Slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
- `ca`: Number of major vessels (0-3) colored by fluoroscopy
- `thal`: Thallium heart scan results (1 = normal; 2 = fixed defect; 3 = reversible defect)
- `target`: Diagnosis of heart disease (1 = heart disease; 2 = no heart disease)

Objectives

In this assignment, you will explore the Cleveland heart disease dataset to answer two personalized research questions. Your specific research questions are determined by the last two digits of your student (or staff) number (see the lookup table below).

Research Question 1 (ANOVA/Kruskal-Wallis): *Does maximum heart rate differ across [Grouping variable] categories?*

Research Question 2 (Chi-Square/Fisher's Exact): *Is the presence of heart disease associated with [Binary predictor]?*

Steps to complete the assignment

Step 1: Create an individualized dataset

Download the Cleveland heart disease dataset (see Downloads section at the end of this document) and follow the steps below to create an individualized dataset with 200 patients:

> **!** Important
>
> You must use your individualized dataset for all analyses in this assignment. Each dataset is uniquely sampled based on your student or staff number, ensuring that every student works independently with a unique dataset. This approach also ensures that results remain reproducible and can be individually verified by the instructor.

- Load the dataset into R
- Set a random seed using your student or staff number
  - Remove any letters (e.g., `S` or `P`) and use the numeric part
  - Example: `set.seed(123456)` for student number `S123456`
- Randomly sample 200 patients from the dataset
  - Use: `heart_data <- heart_data[sample(nrow(heart_data), 200), ]`
- Verify the dataset contains exactly 200 rows
  - Example: `nrow(heart_data)` should return `200`

```
# Load required packages for importing and cleaning the data
# You may load additional packages as needed for your analyses
library(haven)
library(dplyr)
```

```r
# Import the SPSS file
heart_data <- read_sav("datasets/heart_disease_cleveland.sav")

# Convert all categorical variables to factors
heart_data <- heart_data |> mutate(across(where(is.labelled), as_factor))

# Set the random seed
# (replace 123456 with your student or staff number without S or P)
set.seed(123456)

# Create an individualized dataset with 200 patients
heart_data <- heart_data[sample(nrow(heart_data), 200), ]

# Verify the dataset contains exactly 200 rows
nrow(heart_data)
```

Step 2: Identify your personalized research questions

Use the last two digits of your student number to find your personalized research questions in the table below:

Last 2 digits	RQ1: Grouping variable	RQ2: Binary predictor
00–24	Chest pain type (`cp`)	Sex (`sex`)
25–49	Chest pain type (`cp`)	Fasting blood sugar (`fbs`)
50–74	Thalassemia (`thal`)	Sex (`sex`)
75–99	Thalassemia (`thal`)	Fasting blood sugar (`fbs`)

i Example

A student with number **S2734567** has last two digits **67**, which falls in the 50–74 range. Their research questions are:

- **RQ1**: *Does maximum heart rate differ across thalassemia categories?*
- **RQ2**: *Is the presence of heart disease associated with sex?*

Step 3: Create a baseline characteristics table

- Include all variables in the dataset apart from the outcome variable `target`
  - Summarize demographic variables (e.g., `age`, `sex`)

  - Summarize clinical variables (e.g., `chol`, `trestbps`, `thalach`, `cp`)

- Decide on suitable summary measures for each variable
  - Use appropriate measures for continuous variables (e.g., mean, standard deviation, median, interquartile range)
  - Use frequency counts and percentages for categorical variables
- Present your table clearly
  - Use meaningful labels, headings, and clear formatting

Step 4: Perform the analysis for your first research question

Using your grouping variable from the lookup table:

- Visualize the data
  - Create a boxplot to compare the distribution of maximum heart rate across the categories of your grouping variable
- Calculate the estimated population means and 95% confidence intervals for maximum heart rate for each category of the grouping variable
- Select and perform an appropriate test
  - Use one-way ANOVA if normality and equal variances are met
  - Use Kruskal-Wallis test if assumptions are violated
  - Perform post-hoc comparisons using Bonferroni adjusted p-values if significant differences are found

Step 5: Perform the analysis for your second research question

Using the binary predictor from the lookup table:

- Summarize the data
  - Calculate and report the prevalence of heart disease for each category of your binary predictor
  - Include percentages and 95% confidence intervals for each group
- Select and perform an appropriate test
  - Use a Chi-Square test of homogeneity or Fisher's Exact Test, depending on the expected cell counts

Step 6: Write a report

Your report should be structured in the form of **Methods** and **Results** sections, as typically encountered in scientific papers.

- **Methods**
  - State your personalized research questions (based on your student number)
  - Outline the steps taken to analyze the data
  - Describe statistical tests performed, assumptions checked, and adjustments applied
- **Results**
  - Include the baseline characteristics table
  - Present key findings for each research question
  - Include visualizations (e.g., boxplots) to support your findings where applicable
- **Formatting guidelines**
  - Properly label all tables and figures
  - Limit the report to 2–3 pages, including visuals and tables

Reporting examples

When presenting your analysis results, ensure clarity and adherence to proper reporting conventions. Use the following examples as a guide:

> The mean cholesterol levels (95% CI) for the four chest pain types were as follows: typical angina, 245.3 mg/dL (95% CI: 230.1, 260.5); atypical angina, 220.4 mg/dL (95% CI: 205.7, 235.1); non-anginal pain, 230.2 mg/dL (95% CI: 215.6, 244.8); and asymptomatic chest pain, 200.1 mg/dL (95% CI: 185.4, 214.8). A one-way ANOVA was conducted to compare cholesterol levels across these groups, revealing a significant difference, $F(3, 299) = 4.32$, $p = 0.006$. Post-hoc pairwise comparisons using Bonferroni-adjusted p-values indicated that patients with typical angina had significantly higher cholesterol levels compared to those with asymptomatic chest pain (adjusted $p = 0.015$). No other pairwise differences were statistically significant after adjustment.
>
> The prevalence of heart disease was higher among patients older than 65 (68.5%, 95% CI: 60.2%, 76.8%) compared to those 65 or younger (47.2%, 95% CI: 35.6%, 58.8%). Fisher's Exact Test indicated a significant difference between these groups ($p = 0.028$).

Submission instructions

Submit the following files as part of your assignment:

- **The report:** Provide your report in Word or PDF format
- **The R script:** Include your R script (.R file) with all analysis code

Downloads

Cleveland heart disease dataset (SPSS format)