Advanced Medical Statistics

Assignment part 1: Cleveland heart disease dataset

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 the following research questions:

- Does maximum heart rate achieved differ across chest pain types?
- Is the presence of heart disease associated with sex or fasting blood sugar?

Steps to complete the assignment

Step 1: Create an individualized dataset

Download the Cleveland heart disease dataset from Brightspace. After loading the dataset into R or SPSS, 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.

Instructions for R users

- 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`

```r
# Load the required libraries
library(haven)
library(dplyr)

# Load the SPSS file
heart_data <- read_sav("heart_disease_cleveland.sav")
```

```r
# Convert all variables to factors where needed
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)
```

Instructions for SPSS users

- Open the dataset in SPSS
- Open the **Syntax Editor** in SPSS (`File` → `New` → `Syntax`)
- Copy and paste the syntax from the box below into the Syntax Editor, replacing `123456` with your numeric student or staff number without any letters (e.g., `S` or `P`)
- Run the syntax to create a custom dataset with 200 patients
  - Go to `Run` → `All` to run the syntax
- Verify the dataset contains exactly 200 cases
  - Go to `Analyze` → `Descriptive Statistics` → `Frequencies`
  - Select the `sex` variable and click `OK`
  - The output should show `200` cases
- Save the dataset as a new file for further analysis

```
* Set random seed (replace 123456 with your student or staff number without any letters).
SET SEED = 123456.

* Initialize sampling counters for 200 samples from 303 cases.
DO IF $CASENUM = 1.
  COMPUTE #s_$_1 = 200.
  COMPUTE #s_$_2 = 303.
END IF.

* Perform sampling logic.
DO IF #s_$_2 > 0.
  COMPUTE filter_$ = RV.UNIFORM(0, 1) * #s_$_2 < #s_$_1.
  COMPUTE #s_$_1 = #s_$_1 - filter_$.
```

```
  COMPUTE #s_$_2 = #s_$_2 - 1.
ELSE.
  COMPUTE filter_$ = 0.
END IF.

* Apply filter to retain sampled cases.
VARIABLE LABELS filter_$ '200 from the first 303 cases (SAMPLE)'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

! Important note for SPSS users

The syntax above will generate a custom dataset containing 200 patients randomly selected from the original dataset. The selected patients are identified using the filter variable `filter_$`, which is added as a column in the dataset.
To reapply the filter when you reopen the dataset, follow these steps:

1. Go to `Data` → `Select Cases`.
2. Choose `Use filter variable` and select `filter_$` from the list.
3. Click `OK` to apply the filter.

Please note that this filtering step must be repeated each time you open the dataset to ensure that only the selected patients are included in your analysis.

Step 2: 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 3: Perform the analysis for the first research question

Research Question: *Does maximum heart rate achieved differ across chest pain types?*

- Visualize the data
  - Create a boxplot to compare the distribution of maximum heart rate achieved across the four chest pain types
- Calculate the estimated population means and 95% confidence intervals for maximum heart rate achieved for each of the four chest pain types
- 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 4: Perform the analysis for the second research question

Research Question: *Is the presence of heart disease associated with sex or fasting blood sugar?*

- Summarize the data
  - Calculate and report the prevalence of heart disease for each group within sex and fasting blood sugar
  - 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 5: Write a report

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

- **Methods**
  - 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 analysis file:** Include your R script or SPSS output file