Medical Statistics

Lab 5: Cleveland heart disease dataset

Introduction

Welcome to lab 5 in the Medical Statistics course. In this lab, you will combine your learnings from the previous lectures and labs in part I of the course to address 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?

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)

Steps to complete the lab

Step 1: Download and load/open the dataset

Download the Cleveland heart disease dataset (heart_disease_cleveland.sav) from the Datasets menu. SPSS users can simply double click on the downloaded file to open the dataset. R users can follow the instructions below to load and clean the dataset:

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

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

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

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

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