Exercise Sheet 6:
Descriptive Statistics & Linear Regression with Medical Data

Overview

In this exercise sheet we explore descriptive statistics, contingency tables, and simple linear regression using a real medical dataset:

`birthwt` (from the `MASS` package): a study of risk factors for low birth weight in infants.

You will practice:

- Frequency tables and barplots (categorical variables)
- Histograms and ECDF (continuous variables)
- Measures of central tendency and variability
- Boxplots, density and violin plots
- Contingency tables, proportions, chi-square test, odds ratio
- Scatterplots, correlation
- Simple linear regression in a medical context

Unless stated otherwise, you may assume the following R setup:

```
library(MASS)
library(dplyr)
library(ggplot2)
library(forcats)

data(birthwt)
bwt <- birthwt %>%
  mutate(
    race  = factor(race, levels = c(1,2,3),
                   labels = c("White","Black","Other")),
    smoke = factor(smoke, levels = c(0,1),
```

```
                labels = c("No","Yes")),
  ht    = factor(ht,    levels = c(0,1),
                labels = c("No","Yes")),
  ui    = factor(ui,    levels = c(0,1),
                labels = c("No","Yes")),
  low   = factor(low,   levels = c(0,1),
                labels = c("No","Yes"))
)
```

1 Getting to Know the Data

Exercise 1

1. In your own words, describe what one row of the object `bwt` represents.
2. Identify which variables are categorical and which are numeric.
3. For each categorical variable, write down the meaning of its levels.

2 Categorical Variables & Frequency Tables

Exercise 2

1. Create a bar chart for and `low` using `ggplot2`.
2. Briefly comment on the proportion of low birth weight (`low = "Yes"`).

3 Continuous Variables: Histograms & ECDF

Exercise 3

Consider the birth weight variable `bwt` (in grams).

1. Produce histograms of `bwt` with three different bin widths, e.g. 100, 250, and 500.
2. Describe how the choice of bin width changes the appearance of the histogram.
3. Does the distribution of birth weight appear roughly symmetric, right-skewed, or left-skewed?

Exercise 4

1. Plot the empirical cumulative distribution function (ECDF) of `bwt` using `stat_ecdf()`.
2. From the ECDF or directly in R, estimate the proportion of babies with birth weight $\leq 2500\,\text{g}$.

3. Estimate the median birth weight visually from the ECDF and compare with the value from `median(bwt$bwt)`.

4 Central Tendency & Variability

Exercise 5

For the birth weight `bwt`:

1. Compute the six-number summary $(\min, Q_1, \tilde{x}, \bar{x}, Q_3, \max)$.
2. Compare the mean and median. What does their relationship suggest about the skewness of the distribution?

Exercise 6

For the maternal weight at last menstrual period `lwt` (in pounds):

1. Compute the range, interquartile range, and standard deviation of `lwt`.
2. Which measure (IQR or SD) would you prefer as a measure of spread if there are outliers? Explain briefly.
3. Use a boxplot of `lwt` to inspect for potential outliers.

5 Boxplots, Density & Violin Plots

Exercise 7

1. Produce a boxplot of birth weight `bwt` by smoking status `smoke`.
2. Compare the median birth weight between smokers and non-smokers.
3. Compare the IQR in both groups and comment on variability.
4. Do there appear to be more outliers in one group than the other?

Exercise 8

1. Produce a density plot of `bwt` stratified by `smoke`.
2. Describe how the distributions of birth weight differ between smokers and non-smokers (shape, location, spread).

Exercise 9

1. Produce a so-called *violin plot* of `bwt` by `smoke`.
2. Compare the violin plot with the boxplot. What additional information does the violin plot reveal about the shape of the distributions?

6 Bivariate Relationships: Scatterplots & Correlation

Exercise 10

Consider the relationship between maternal weight `lwt` and birth weight `bwt`.

1. Make a scatterplot of `bwt` versus `lwt`.
2. Does higher maternal weight appear to be associated with higher birth weight?
3. Are there any obvious outliers or unusual points?

Exercise 11

1. Compute the correlation between `lwt` and `bwt`.
2. Interpret the sign and magnitude of the correlation (direction and strength).
3. Is your interpretation consistent with the scatterplot from Exercise 10?

7 Contingency Tables & Association

Exercise 12

Consider the relationship between smoking status `smoke` and low birth weight indicator `low`.

1. Construct the contingency table `table(bwt$smoke, bwt$low)`.
2. Compute row-wise proportions using `prop.table(..., margin = 1)`.
3. From the table and proportions, which group (smokers or non-smokers) appears to have a higher proportion of low birth weight?

Exercise 13

1. Perform a chi-square test of independence between `smoke` and `low`.
2. State the null and alternative hypotheses.

Exercise 14 (Advanced)

1. Using `fisher.test(tab)`, compute the odds ratio for low birth weight comparing smokers and non-smokers.
2. Interpret the odds ratio: is the odds of low birth weight higher for smokers? By approximately what factor?

8 Simple Linear Regression

Exercise 15

Fit a simple linear regression model for birth weight `bwt` as a function of maternal weight `lwt`:

1. Fit the model `lm(bwt ~ lwt, data = bwt)` and inspect the summary.
2. Write down the estimated regression equation.
3. Interpret the slope coefficient in words: what does a one-pound increase in maternal weight predict for birth weight?

Exercise 16

1. Add the regression line to the scatterplot of `lwt` vs. `bwt` (using `geom_smooth(method = "lm")`).
2. Visually assess whether a straight line seems to fit the data reasonably well.
3. Suggest one additional predictor from `bwt` (e.g. `smoke`, `age`, `ht`) that might improve prediction of birth weight.

9 Group Comparisons: Smoke vs No Smoke

Exercise 17

Using grouped summary statistics:

1. Compute the mean and median birth weight separately for smokers and non-smokers.
2. Compare these values.
3. Summarize in 3–4 sentences how smoking appears to be related to birth weight in this dataset, referencing at least two of the following: boxplots, density/violin plots, group means/medians, or contingency tables.

10 Summary Reflection

Exercise 18

1. Explain why it is important to look at both numerical summaries (e.g. mean, standard deviation, correlation) and graphical summaries (e.g. histograms, boxplots, scatterplots) when analyzing data.

2. Give one specific example from the preceding exercises where numerical summaries alone were insufficient, and a plot or contingency table provided crucial additional information.