Medical Statistics – Lab 1

R version

Welcome to lab 1 in the medical statistics course. In this lab, we will explore descriptive statistics and probability calculations for random variables. We will use an example dataset to practice summarizing continuous and categorical variables, and introduce some basic concepts of probability distributions.

> **Learning R**
>
> Throughout this course, we use the `ggplot2` package for data visualization and the `dplyr` package for data transformation and summary. Students who want to learn more about these packages are recommended to study chapters 1 (Data visualization) and 3 (Data transformation) from Hadley Wickham's *R for Data Science* book: https://r4ds.hadley.nz/

Part 1 - Descriptive statistics

For this part of the lab, we will use the `lowbwt.sav` dataset. This dataset contains information collected at Baystate Medical Center Springfield, MA, during 1986 and is from Appendix 1 of Hosmer and Lemeshow (1989). The data include information about factors related to low birth weight, an important concern due to its association with high infant mortality rates and birth defects.

The dataset is in SPSS format, but you can easily load it into R using the `haven` package. To get started, make sure to install the necessary packages:

```
install.packages("haven")
install.packages("dplyr")
install.packages("ggplot2")
```

Then, load the dataset as follows:

```r
library(haven)
library(dplyr)
library(ggplot2)

# Load the dataset
lowbwt <- read_sav("datasets/lowbwt.sav")

# Convert alll labelled variables into factor variables
lowbwt <- lowbwt |> mutate(across(where(is.labelled), as_factor))
```

The dataset includes the following variables:

Variable	Abbreviation
Identification Code	ID
Low Birth Weight (0 = Birth Weight 2500g, 1 = Birth Weight < 2500g)	low
Age of the Mother in Years	age
Weight in Pounds at the Last Menstrual Period	lwt
Ethnicity (1 = Caucasian, 2 = Afro-American, 3 = Asian)	ethnicity
Smoking Status During Pregnancy (1 = Yes, 0 = No)	smoke
History of Premature Labor (0 = None, 1 = One, etc.)	ptl
History of Hypertension (1 = Yes, 0 = No)	ht
Presence of Uterine Irritability (1 = Yes, 0 = No)	urirr
Number of Physician Visits During the First Trimester (0 = None, 1 = One, 2 = Two, etc.)	pvft
Birth Weight in Grams	bwt

The variables `bwt` (birth weight in grams) and `low` (low birth weight) are the outcome variables, while all other variables are considered independent variables. The outcome variables will be analyzed in more detail in subsequent labs. In this lab, we will focus on describing the study population in terms of the independent variables to obtain a set of summary statistics that could be used to populate a baseline characteristics table.

Descriptive analysis of continuous variables

Let's start by calculating the summary statistics for the continuous variable `age`. We'll use the `summarise()` function from the `dplyr` package, which allows us to calculate multiple summary statistics in a single, organized call:

```
age_summary <- lowbwt |>
  summarise(
    mean = mean(age, na.rm = TRUE),
    sd = sd(age, na.rm = TRUE),
    median = median(age, na.rm = TRUE),
    q1 = quantile(age, 0.25, na.rm = TRUE),
    q3 = quantile(age, 0.75, na.rm = TRUE),
    iqr = IQR(age, na.rm = TRUE)
  )

age_summary
```

The `summarise()` function creates a new data frame containing the summary statistics you specify. The functions `mean()`, `sd()`, `median()`, and `IQR()` are built-in base R functions that work within `summarise()`. The `quantile()` function calculates the quartiles: Q1 (25th percentile) and Q3 (75th percentile). The `na.rm = TRUE` argument ensures that any missing values are ignored in the calculations.

> **i** IQR vs [Q1, Q3]
>
> Note the difference between IQR and the quartile interval:
>
> - **IQR** (Interquartile Range) = Q3 - Q1, which is a single number representing the spread of the middle 50% of the data
> - **[Q1, Q3]** is the interval itself, showing the actual range where the middle 50% of observations fall
>
> In scientific papers, it's more common to report the median along with [Q1, Q3] (e.g., "median age: 23 years [19, 26]") rather than reporting the IQR as a single value.

To decide which summary measures (mean and standard deviation, or median and [Q1, Q3]) are appropriate to report, we need to understand the shape of the distribution of the `age` variable. We do this by creating a histogram:

```
ggplot(lowbwt, aes(x = age)) +
  geom_histogram(binwidth = 2,
```

```
                 fill = "blue",
                 color = "black") +
  labs(title = "Histogram of Age",
       x = "Age of Mother (years)",
       y = "Frequency") +
  theme_minimal()
```

> Question 1
>
> Based on the shape of the histogram, determine which summary statistics are more appropriate to report.

> Question 2
>
> Calculate the mean, standard deviation, median, Q1, Q3, and IQR for the variable `lwt` using the `summarise()` function. Additionally, create a histogram to determine the shape of its distribution and decide which summary measures are most appropriate to report.

Descriptive analysis of categorical variables

Let's move on to analyzing the categorical variables. We'll calculate the frequency and percentage of mothers who smoked during pregnancy (`smoke`) using the `count()` and `mutate()` functions from `dplyr` package:

```
smoke_summary <- lowbwt |>
  count(smoke) |>
  mutate(percentage = n / sum(n) * 100)

smoke_summary
```

The `count()` function calculates the frequency of each category and automatically creates a new column called `n` that contains these frequencies. The `mutate()` function then adds another column with the percentages, calculated by dividing each frequency (`n`) by the total number of observations (`sum(n)`) and multiplying by 100.

> Question 3
>
> Calculate the frequencies and percentages for the variable history of hypertension (`ht`).

In addition to calculating frequencies and percentages, it can also be helpful to visualize categorical data. One common way to do this is by creating a bar chart. Below is an example of how you can create a bar chart for the `smoke` variable:

```
ggplot(lowbwt, aes(x = factor(smoke))) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Smoking Status During Pregnancy",
       x = "Smoking Status",
       y = "Frequency") +
  theme_minimal()
```

This bar chart shows the frequency of mothers who smoked and those who did not during pregnancy. Similarly, you can create a bar chart for the variable `ht` to visualize the frequency of mothers with a history of hypertension.

> Question 4
>
> Create a bar chart for the variable `ht` to visualize the frequency of mothers with a history of hypertension.

Part 2 - Probability calculations for random variables

In this section, we will focus on two commonly used probability distributions: the binomial distribution and the normal distribution. We will explore how to work with these distributions in R using practical examples.

Binomial Distribution

A binomial distribution represents the number of successes in a fixed number of independent trials, each with the same probability of success. For example, if we have 10 patients and we want to know the probability that exactly 3 of them respond to a given treatment, where the response rate is known to be 40%, we can use the `dbinom()` function:

```
# Probability that exactly 3 out of 10 patients respond
# to the treatment (assuming p = 0.4)
p_response_3 <- dbinom(3, size = 10, prob = 0.4)
p_response_3
```

Explanation of `dbinom()` Arguments

- **`x`**: The number of successes we are interested in (in this example, `3` patients responding).
- **`size`**: The number of trials, which represents the total number of patients (in this example, `10` patients).
- **`prob`**: The probability of success in each trial (in this example, `0.4` or 40% response rate).

To calculate cumulative probabilities, use the `pbinom()` function. For example, to calculate the probability that 3 or fewer patients out of 10 respond to the treatment:

```
p_cum_3 <- pbinom(3, size = 10, prob = 0.4)
p_cum_3
```

Explanation of `pbinom()` Arguments

- **`q`**: The number of successes we want to calculate the cumulative probability for (in this example, `3` or fewer successes).
- **`size`**: The number of trials, which represents the total number of patients (in this example, `10` patients).
- **`prob`**: The probability of success in each trial (in this example, `0.4` or 40% response rate).

By default, the `pbinom()` function calculates the probability that the number of successes is less than or equal to `q`. However, it is also possible to calculate the probability that the number of successes is greater than `q` by setting `lower.tail = FALSE`. The `lower.tail` argument specifies whether the cumulative probability is calculated from the lower tail (default, `TRUE`) or the upper tail (`FALSE`).

Question 5

The probability of being blood group B is 0.08. What is the probability that if 500 ml of blood is taken from each of 100 unrelated blood donors fewer than 1,500 ml of group B blood will be obtained?

Question 6

In a clinical trial in which a total of 100 patients are allocated to two treatments A and B by simple randomization (tossing a coin for each new patient). What is the probability that the difference between the numbers of patients in the two treatment groups exceeds 20? (Hint: the number of individuals in one treatment group (for example A) follows a Binomial distribution).

Normal Distribution

Suppose that we want to calculate the probability that a randomly selected individual has a weight less than or equal to 80 kg, assuming that the distribution of weight in the population follows a normal distribution with mean 72 kg and standard deviation 10 kg.

To calculate this probability, we first need to **standardize** the value using the formula:

$$Z = \frac{x - \mu}{\sigma} = \frac{80 - 72}{10} = 0.8$$

where:

- `x` is the value we want to standardize (in this case, 80 kg).
- is the mean of the distribution (in this case, 72 kg).
- is the standard deviation of the distribution (in this case, 10 kg).

Using R, we can then use the `pnorm()` function to find the corresponding cumulative probability from the standard normal distribution:

```
# Standardize the value
z_value <- (80 - 72) / 10
z_value

# Use the standard normal distribution to find
# the cumulative probability
cum_prob_weight <- pnorm(z_value)
cum_prob_weight
```

Question 7

Over a 25 year period the mean height of adult males increased from 175.8 cm to 179.1 cm, but the standard deviation stayed at 5.84 cm. The minimum height requirement for men to join the police force is 172 cm. What proportion of men would be too short to become policemen at the beginning and end of the 25 year period, assuming that the height of adult males has a Normal distribution?

Part 3 - Some conceptual questions

Question 8

Imagine we have some observations on blood pressure and calculate the mean, standard deviation, median and IQR. How do these measures change if all observations are

a) increased by 10
b) decreased by their mean
c) multiplied by 10
d) divided by their standard deviation

Question 9

Two fair dice are rolled, the results are X1 and X2.

(a) What is the probability Prob(X1=X2)?
(b) What is the expected value E(X1), and the standard deviation SD(X1)?
(c) Give the expected value and the variance of X1+X2 and of X1-X2.

Question 10

There are two hospitals in town. On average 45 deliveries take place each day in the larger hospital, and 15 in the smaller. The probability of a baby being a boy is about 0.52, and the probability of twins is about 0.012. On any day, which hospital is more likely

(a) to have a set of twins delivered
(b) to have more than 60 % of babies being boys?

Question 11

The probability of a baby being a boy is 0.52. For six women delivering consecutively in the same labour ward on one day, which of the following exact sequences of boys and girls is most likely and which least likely?
GBGBGB
BBBGGG
GBBBBB