Advanced medical Statistics – Lab 1

SPSS version

Welcome to lab 1 in the advanced 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.

Part 1 - Descriptive statistics

For this part of the lab, we will use the dataset `lowbwt.sav`, which you can download from Brightspace. 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 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
Socio-Economic Status (1 = Low, 2 = Middle, 3 = High)	ses
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

Variable	Abbreviation
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`.

1. **Mean** and **Standard Deviation**: Use the "Analyze > Descriptive Statistics > Descriptives…" menu in SPSS to calculate the mean age of the mothers. Select `age` as the variable, and ensure that the mean and standard deviation are checked in the "Options" dialog.
2. **Median and Interquartile Range (IQR)**: To calculate the median and IQR for `age`, use the "Analyze > Descriptive Statistics > Frequencies…" menu. Select `age` as the variable and click on "Statistics…" to choose the median and quartiles.

To decide which summary measures (mean and standard deviation, or median and IQR) are appropriate to report, we need to understand the shape of the distribution of the `age` variable. Create a histogram by using "Graphs > Legacy Dialogs > Histogram…" and selecting `age` as the variable.

> 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, and IQR for the variable `lwt`. 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 will start by calculating the frequency and percentage of mothers who smoked during pregnancy (`smoke`):

1. **Frequency Table**: Use the "Analyze > Descriptive Statistics > Frequencies…" menu to calculate the frequency of each category for `smoke`.
2. **Percentage Calculation**: SPSS will automatically calculate the percentages for each category in the frequency table output.

Question 3

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

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. To create a bar chart for the `smoke` variable, use "Graphs > Legacy Dialogs > Bar…" and select "Simple" and "Summaries for Groups of Cases." Then select `smoke` as the Category Axis.

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 SPSS 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 SPSS to calculate this probability.

To perform this calculation in SPSS, go to "Transform > Compute Variable…" and create a new variable called `p_response_3`. Use the function `PDF.BINOM(x, n, p)` to calculate the cumulative probability:

- `x`: The number of successes we are interested in (in this example, `3` patients responding).
- `n`: The number of trials (in this example, `10` patients).
- `p`: The probability of success in each trial (in this example, `0.4` or 40% response rate).

Cumulative probabilities, such as the probability that 3 or fewer patients out of 10 respond to the treatment, can similarly be calculated by creating a new variable with the `CDF.BINOM(x, n, p)` function.

Note: Unlike other statistical software, SPSS requires that calculations performed using the "Transform > Compute Variable…" tool always have a target variable specified. This means you cannot perform calculations without creating an output variable in your dataset, which can make one-time calculations cumbersome. This is a limitation of SPSS, especially when you just want to explore different scenarios without cluttering your dataset with temporary variables.

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 SPSS, we can then use the "Transform > Compute Variable..." menu and the `CDF.NORMAL(z, mean, sd)` function to find the corresponding cumulative probability from the standard normal distribution by setting `z = 0.8`, `mean = 0` and `sd = 1`.

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