Data 1204: Statistical and Predictive Modelling

Assignment #2

Hypothesis Testing

DURHAM COLLEGE

SUCCESS MATTERS

Name: **Raj Dholakia**

Student Number: **100813041**

Lecturer(s): **Ritwick Dutta**

Table of Contents

1 Introduction

In this assignment, we are looking at a small blood pressure dataset consisting of 25 data points. The dataset tries to better understand the impact of regular exercise on blood pressure of the participants. The experiment was carried out on 25 individuals, hence, the **25 data points**. The dataset has **three columns (features)**:

1. **Subject**: this is an identification number for the subjects that were part of the experiment. The variable **ranges from 1 to 25.**
2. **Before**: this is the **blood pressure of the subject** at the beginning of the experiment.
3. **After**: this is the **blood pressure of the subject after thirty (30) days** into the experiment. These thirty (30) days the participant is expected to exercise on a regular basis.

The goal of the experiment, as defined in the assignment, is to determine the effect of regular exercise on the blood pressure of the subject.

A null hypothesis testing will be carried out to check the results of the experiment. The results obtained from the experiment are saved in **BloodPressure.xls**.

2 Hypothesis Statement

A null hypothesis can be defined by assuming exercising has no impact on blood pressure. Hence, we expect no change in blood pressure after thirty (30) days of regular exercise.

<u>H_0 (null hypothesis): Regular exercise does **not** have any impact on blood pressure.</u>
Mathematically: **difference in blood pressure before and after should be equal to 0**.

<u>H_a (Alternate hypothesis): Regular exercise brings changes in blood pressure.</u>

3 Process to Test the Hypothesis

The following steps will be carried out to check if the null hypothesis is true or false.

1. Find the **absolute percentage change (percent_change)** in blood pressure of each participant before and after the experiment. Here we take the absolute difference to get rid of biases due to the direction of change (it can be negative or positive).

$$\boldsymbol{percent\_change = abs\left(\frac{BP_{before} - BP_{after}}{BP_{before}}\right)}$$

   If the null hypothesis is true, we expect mean of percent_change to be close to 0.

2. In this step, we **calculate the mean and standard error** in the available twenty-five (25) data points. As we are using absolute percentage change, the mean cannot average out zero (0) due to opposite signs. For example, consider the following set of difference:

   {-10, -15, -20, 0, 20, 15, 10}

   In the above set, some of the changes in blood pressure are negative, while some are positive. Though from the numbers, it is evident that there are differences between before and after, the mean of the set is zero (0).
   This gives a false indication most of the values being 0. When in reality, on an average there is an absolute change of approximately fifteen (15). Hence, the mean of differences becomes more reliable when using the absolute value. Furthermore, using percentage values standardizes the changes between 1 and 100%.

   *Note: As the sample size is **less than thirty (30)**, we will be using the t-test and **not** the z-test to test the null hypothesis. (Glen, n.d.).*

3. We use the mean of **zero (0)** and compare it to the **mean of *percent_change***. We calculate the **p-value** using the **one-tail test**. Using the

absolute differences means we cannot have negative values for ***percent_change***, hence we use the one-tail test to see if the mean of values is greater than a certain value.

4. If the p-value found is **less than 0.05 or 5%**, the null hypothesis can be assumed false as we can say that there is a 95% probability that regular exercise **does** bring about a change in blood pressure. If the value is greater than 0.05, the null hypothesis **cannot** be dis-proved and **nothing can be concluded**.

4 Testing the Hypothesis in R

4.1 About the notebook

In this notebook, we will test the following hypothesis:

$$H_0: Regular\ exercise\ does\ not\ have\ any\ impact\ on\ blood\ pressure.$$

The process followed to determine if the null hypothesis is true is described in the word document. As the sample size is less than thirty (30), the t-distribution table is used to determine the p-value.

4.2 Preparing the Environment and Loading Data

Install required packages.

```
# install.packages(c("lattice", "readr", "readxl"))
```

Loading the required libraries.

```
library(lattice)
library(readxl)
library(readr)

options(digits=3)
```

Loading blood pressure data.

```
bp_data <- read_excel("BloodPressure.xls")
bp_data

## # A tibble: 25 x 3
##    Subject Before After
##      <dbl>  <dbl> <dbl>
##  1       1    135   127
##  2       2    142   145
##  3       3    137   131
##  4       4    122   125
##  5       5    147   132
##  6       6    151   147
##  7       7    131   119
##  8       8    117   125
##  9       9    154   132
## 10      10    143   139
## # … with 15 more rows
```

4.3 Calculating Key Statistics

Calculating the absolute percentage change in blood pressure of the subjects before and after the 30-day experiment.

```
bp_data$percent_change <- abs((bp_data$After - bp_data$Before)/bp_data$Bef
ore)
bp_data

## # A tibble: 25 x 4
##    Subject Before After percent_change
##      <dbl>  <dbl> <dbl>          <dbl>
##  1       1    135   127         0.0593
##  2       2    142   145         0.0211
##  3       3    137   131         0.0438
##  4       4    122   125         0.0246
##  5       5    147   132         0.102
##  6       6    151   147         0.0265
##  7       7    131   119         0.0916
```

```
## 8          8     117    125           0.0684
## 9          9     154    132           0.143
## 10        10     143    139           0.0280
## # … with 15 more rows
```

Calculating the mean, standard deviation and mean standard error of the sample provided. This will be used to calculate the t-value, which will further be used to calculate the p-value.

```
mean.percent_change <- mean(bp_data$percent_change)
sd.percent_change <- sd(bp_data$percent_change)
SE.percent_change <- sd.percent_change / sqrt(length(bp_data$percent_chang
e))
mean.percent_change

## [1] 0.0677

sd.percent_change

## [1] 0.0547

SE.percent_change

## [1] 0.0109
```

4.4 Testing the Null Hypothesis (p-value calculation): One-tail Test

Null hypothesis:

$$H_0 = 0$$

Alternate hypothesis:

$$H_a > 0$$

Calculating the t-value.

```
# State the Ho value and calculate the t-score
Ho <- 0
```

```
t <- (mean.percent_change - Ho) / SE.percent_change
t

## [1] 6.19
```

Calculating the p-value.

```
p_value <- pt(-t, df=length(bp_data$percent_change)-1)
p_value

## [1] 1.06e-06

p_value < 0.05 # Check if the p-value is less than 0.05 or 5%.

## [1] TRUE
```

5 Conclusion

From the above analysis, the p-value was found to be **1.06e-06**. This is way smaller than 0.05, which is the threshold determined for the hypothesis test at the beginning of the analysis. Hence, it can be confidently concluded that the **null hypothesis is false**. It can be rejected and the alternate hypothesis, which states that there is an impact on the blood pressure due to regular exercise, can be deemed true. **Thirty (30) days of regular exercise does bring changes to blood pressure.**

6 References

Glen, S. (n.d.). T-Distribution Table (One Tail and Two-Tails). Retrieved February 7, 2021, from From StatisticsHowTo.com: Elementary Statistics for the rest of us! website: https://www.statisticshowto.com/tables/t-distribution-table/