Computational Statistics and Probability

PROJECT 1 - Report

- Andria Grace

TABLE OF CONTENTS

Introduction:

Hypothesis testing is a fundamental statistical technique used to assess the validity of a proposed hypothesis by comparing it to observed data. In this report, we will discuss the key components of hypothesis testing, including the formulation of null and research hypotheses, and the interpretation of results.

Formulation of Null and Research Hypotheses:

The first step in hypothesis testing is to formulate two hypotheses: the null hypothesis (H0) and the research hypothesis (H1).

Null Hypothesis (H0):
The null hypothesis represents the default or status quo assumption. It states that there is no significant effect, no difference, or no relationship between the variables being studied. In a mathematical notation, H0 typically contains an equal sign,
e.g., H0: $\mu = 50$ (which asserts that the population mean is equal to 50).

Research Hypothesis (H1):
The research hypothesis, also known as the alternative hypothesis, contradicts the null hypothesis. It asserts that there is a significant effect, difference, or relationship between the variables. H1 does not contain an equal sign and may take different forms, depending on the nature of the research question.
For example, H1: $\mu \neq 50$ (which asserts that the population mean is not equal to 50).

Conducting the Hypothesis Test:
Once the null and research hypotheses are formulated, a statistical test is conducted using sample data to determine whether there is enough evidence to reject the null hypothesis in favor of the research hypothesis. The choice of the appropriate statistical test depends on the research question and the data type
(e.g., t-test, chi-square test, ANOVA, etc.).

Objective:

This study aims to investigate potential disparities in systolic and diastolic blood pressure among different groups categorized by physical activity levels (high activity vs. low activity) and BMI (high BMI vs. normal BMI).

Computational Statistics and Probability

Statement of the problem

1. Load the NHANES dataset from the Final Project (as per instructions provided).
2. Create a new object with age, biological sex, height, weight, systolic and diastolic blood pressure, diabetes, and physical activity. Recode all categorical variables into factors.
3. Compute descriptive statistics for all included variables for the full dataset (include in report as Table 1).

Case scenario:
An investigator wants to study the effects of
a) high and low levels of physical
activity and
b) high and normal BMI (please use 18.5 to 25 kg/m^2 for normal and >30 kg/m^2 as high), on systolic and diastolic blood pressure.

For the purpose of hypothesis generation for a prospective study, the investigator wants to compare, notably disregarding the problem of multiple testing, random samples (of 250 individual participants) within these groups to the overall population mean of systolic and diastolic blood pressure. Provide analytic support to the researcher.

1. Create tables with descriptive statistics (Table 2a and 2b), histograms (Figure 1a to d) and do formal hypothesis testing using two-sided and one-sided z tests.

Step 1:

Data Loading and Preprocessing

a) **Load Required Libraries:**

library(NHANES): Loads the NHANES package, which contains datasets and functions related to the National Health and Nutrition Examination Survey (NHANES) data.

b) **Load and Clean NHANES Data.**

data("NHANES"): Loads the NHANES dataset into your R session.

NHANES <- NHANES[!duplicated(NHANES$ID),]: Removes duplicate records in the NHANES dataset based on the "ID" variable.

c) **Load the nhanesA Package:**

library(nhanesA): Loads the nhanesA package, which provides additional functionalities working with NHANES data.

d) **Load NHANES Tables:**

The **nhanesTables** function is used to load specific NHANES tables (DEMOGRAPHICS, DIETARY, EXAMINATION, LABORATORY, QUESTIONNAIRE) for the year 2010.

Step 2:

Data Selection and Handling

e) **Select Variables of Interest:**

Creates a new object **selected_data** by selecting specific variables of interest, including Age, Gender, Height, Weight, BPSysAve (Systolic Blood Pressure), BPDiaAve (Diastolic Blood Pressure), Diabetes, PhysActive, and BMI from the NHANES dataset.

f) **Check Data Types:**

Uses the **class** function to check the data types of various variables within **selected_data**. This step confirms that categorical variables are already in factor format.

g) **Replace Missing Values:**

- Defines a function **replace_na_with_mode** to replace missing values (NAs) in categorical columns with the mode (most frequent value). The function iterates through columns, identifies categorical columns, and replaces NAs with the mode.

- Calls the **replace_na_with_mode** function to handle missing values in categorical columns.

h) **Replace Missing Values with Column Averages:**

- Replaces missing values in numeric columns (Age, Height, Weight, BMI, Systolic BP, and Diastolic BP) with the mean (average) value of the respective column, calculated using the mean function.

Step 3:

Descriptive Statistics

i) Check the Structure of the Modified Data:

Uses **str** to check the structure of the modified **selected_data**, which includes variable types and some sample data.

j) Summary Statistics (Table 1):

- Generates a summary of descriptive statistics for all included variables in **selected_data** using the **summary** function.
- Prints "Table 1: Descriptive Statistics for All Included Variables."

```
print(table1)
```

```
##       Age           Gender         Height          Weight          BPSysAve
##  Min.   : 0.00   female:3420   Min.   : 83.6   Min.   :  2.80   Min.   : 76
##  1st Qu.:15.00   male  :3359   1st Qu.:156.0   1st Qu.: 53.60   1st Qu.:108
##  Median :34.00                 Median :164.3   Median : 70.80   Median :118
##  Mean   :35.45                 Mean   :160.4   Mean   : 69.06   Mean   :118
##  3rd Qu.:54.00                 3rd Qu.:173.2   3rd Qu.: 87.40   3rd Qu.:124
##  Max.   :80.00                 Max.   :200.4   Max.   :230.70   Max.   :226
##     BPDiaAve      Diabetes   PhysActive      BMI
##  Min.   :  0.00   No :6227   No :2473   Min.   :12.88
##  1st Qu.: 62.00   Yes: 552   Yes:4306   1st Qu.:21.50
##  Median : 66.72                         Median :26.21
##  Mean   : 66.72                         Mean   :26.49
##  3rd Qu.: 74.00                         3rd Qu.:30.34
##  Max.   :116.00                         Max.   :81.25
```

This table provides summary statistics for various variables.

- For the "Age" variable, it includes the minimum, 1st quartile, median, mean, 3rd quartile, and maximum values.
- "Gender" shows the counts of "female" and "male."
- Numeric variables like "Height," "Weight," "BPSysAve," "BPDiaAve," and "BMI" have minimum, 1st quartile, median, mean, 3rd quartile, and maximum values.
- Categorical variables "Diabetes" and "PhysActive" are summarized by displaying counts of "No" and "Yes" categories.

Step 4:

Hypothesis Testing

k) Define Z-Test and Calculation Functions:

- Functions for conducting z-tests and calculating statistics are defined.
- Descriptive statistics for systolic and diastolic blood pressure are calculated for high and low physical activity levels and high and normal BMI categories.
- Tables 2a and 2b are created, showing descriptive statistics for systolic and diastolic blood pressure, respectively, for the specified groups.

```
print(table_2a)
```

```
##           Group Mean_Systolic_BP SD_Systolic_BP Min_Systolic_BP Max_Systolic_BP
## 1 High Activity         115.9443       14.33607              79             210
## 2  Low Activity         121.5322       18.11770              76             226
## 3      High BMI         122.3050       16.47776              76             226
## 4    Normal BMI         115.1441       16.89881              78             221
```

Table 2a: Descriptive Statistics for Systolic Blood Pressure:

- This table presents descriptive statistics for systolic blood pressure within four groups: "High Activity," "Low Activity," "High BMI," and "Normal BMI."
- The statistics include mean, standard deviation, minimum, and maximum values for systolic blood pressure within each group.

```
print(table_2b)
```

```
##           Group Mean_Diastolic_BP SD_Diastolic_BP Min_Diastolic_BP
## 1 High Activity          65.54898        13.54243                0
## 2  Low Activity          68.75643        13.33631                0
## 3      High BMI          70.20863        14.02415                0
## 4    Normal BMI          64.98100        13.22948                0
##   Max_Diastolic_BP
## 1              110
## 2              116
## 3              116
## 4              116
```

Similar to Table 2a, this table provides descriptive statistics for diastolic blood pressure within the same four groups: "High Activity," "Low Activity," "High BMI," and "Normal BMI."
The statistics include mean, standard deviation, minimum, and maximum values for diastolic blood pressure within each group.

Step 5:

Hypothesis Testing with Z-Tests

- Random samples for systolic and diastolic blood pressure are generated within specified groups.
- Z-scores are calculated for these samples using the z-test function.
- Hypothesis testing is conducted using two-sided and one-sided z-tests.
- Sets a significance level alpha for hypothesis testing.
- Computes two-sided and one-sided p-values for hypothesis tests comparing the specified groups' blood pressure samples to the population mean. Z-scores are used for the calculations.

```
print(results_systolic_two_sided)
```

```
##           Group P_Value_Systolic_Two_Sided
## 1 High Activity               2.812259e-03
## 2  Low Activity               2.463742e-04
## 3      High BMI               1.494048e-04
## 4    Normal BMI               7.848196e-05
```

```
print(results_diastolic_two_sided)
```

```
##           Group P_Value_Diastolic_Two_Sided
## 1 High Activity                6.194934e-03
## 2  Low Activity                3.416932e-04
## 3      High BMI                5.281923e-06
## 4    Normal BMI                2.033980e-02
```

Both results_systolic_two_sided and results_diastolic_two_sided data frames provide a summary of two-sided hypothesis testing results for systolic and diastolic blood pressure in four groups, respectively i.e "High Activity," "Low Activity," "High BMI," and "Normal BMI." . The P-values in their respective columns allow for the evaluation of statistical significance in differences between the group means and the overall population means.

```
print(results_systolic_one_sided)
```

```
##           Group P_Value_Systolic_One_Sided
## 1 High Activity                 0.9985939
## 2  Low Activity                 0.9998768
## 3      High BMI                -2.7920567
## 4    Normal BMI                 0.9999608
```

```
print(results_diastolic_one_sided)
```

```
##           Group P_Value_Diastolic_One_Sided
## 1 High Activity                9.969025e-01
## 2  Low Activity                9.998292e-01
## 3      High BMI                2.640962e-06
## 4    Normal BMI                9.898301e-01
```

The results_systolic_one_sided and results_diastolic_one_sided data frames contain one-sided (upper tail) hypothesis testing results for systolic and diastolic blood pressure in four groups: "High Activity," "Low Activity," "High BMI," and "Normal BMI." The respective P_Value columns provide p-values for each group comparison, facilitating the assessment of statistical significance in differences between their means in the specified direction.

Step 7:

Print Hypothesis Testing Results:

Prints the results of hypothesis testing for systolic and diastolic blood pressure, both two-sided and one-sided tests, for the specified groups.

Step 8:

Plot Histograms:

Plots histograms for systolic and diastolic blood pressure within the specified groups and for different BMI categories.

These steps involve data loading, cleaning, manipulation, descriptive statistics, hypothesis testing, and data visualization related to NHANES data for blood pressure and BMI analysis.

Histograms for systolic and diastolic blood pressure within "High Activity" and "Low Activity" groups, visualizing the distribution of blood pressure values. The 2x2 layout of histograms allows for a visual comparison of these distributions and the assessment of any differences or similarities between the two activity groups.

Computational Statistics and Probability

Methodology:

Utilizing z-tests, we compared blood pressure values across groups. Two-sided tests and one-sided tests were conducted with a significance level (α) set at 0.05.

Case 1:

Null Hypothesis (H0):
There is no significant difference in systolic blood pressure between individuals with high physical activity levels
Research Hypothesis (H1):
There exists a significant difference in systolic blood pressure based on high physical activity levels
Results:High Activity
Systolic Blood Pressure: p-value_two_sided = 0.002812259. < 0.05
Systolic Blood Pressure: p-value_one_sided = 0.9985939 > 0.05

As two - sided test is following null hypothesis and one sided test is rejecting null hypothesis so we can't reject null hypothesis

Case 2:

Null Hypothesis (H0):
There is no significant difference in diastolic blood pressure between individuals with high physical activity levels
Research Hypothesis (H1):
There exists a significant difference in diastolic blood pressure based on high physical activity level
Diastolic Blood Pressure: p-value_two_sided = 0.006194934 <0.05
Diastolic Blood Pressure: p-value_one_sided = 0.9969025 >0.05

As two - sided test is following null hypothesis and one sided test is rejecting null hypothesis so we can't reject null hypothesis

Case 3:

Null Hypothesis (H0):
There is no significant difference in systolic blood pressure between individuals with low physical activity levels

Research Hypothesis (H1):
There exists a significant difference in systolic blood pressure based on low physical activity levels

Low Activity:
Systolic Blood Pressure: p-value_two_sided = 0.0002463742<0.05
Systolic Blood Pressure: p-value_one_sided = 0.9998768 >0.05

As two - sided test is rejecting null hypothesis and one sided test is following null hypothesis so we can't reject null hypothesis

Case 4:

Null Hypothesis (H0):
There is no significant difference in diastolic blood pressure between individuals with low physical activity levels
Research Hypothesis (H1):
There exists a significant difference in diastolic blood pressure based on low physical activity levels
Diastolic Blood Pressure: p-value_two_sided = 0.0003416932 <0.05
Diastolic Blood Pressure: p-value_one_sided = 0.9998292 >0.05

As two - sided test is rejecting null hypothesis and one sided test is following null hypothesis so we can't reject null hypothesis

CONCLUSION:

This code offers an extensive examination of NHANES data, shedding light on the connections between demographic variables, physical activity, blood pressure, and BMI. Through descriptive statistics, hypothesis testing, and visual aids, it enhances our comprehension of the health characteristics among diverse population subgroups. This analysis holds the potential to identify potential risk factors and guide public health decisions related to cardiovascular health and obesity, providing a comprehensive exploration of the dataset's utility in health research.

The results indicate that there is either significant or insignificant evidence to reject the null hypothesis, with specific groups demonstrating notable differences, if any. These outcomes underscore the pivotal role of physical activity and BMI in deciphering variations in blood pressure. This investigation not only furnishes statistical support for disparities in blood pressure but also underscores the importance of lifestyle factors in cardiovascular health. Please replace "[insert value here]" with the actual p-values derived from your analysis for a complete and precise report.