Module 3: R Practice Assignment (Week 3)

Mohammed Saif Wasay (002815958)

Masters of Professional Studies in Analytics, Northeastern University

ALY 6010: Probability Theory and Introductory Statistics

Harpreet Sharma

June 09th, 2024

Introduction

In this practice project, we will look at the Body Mass Index (BMI) dataset. This data collection has 500 rows and four columns with information about the individual such as gender, height in cm, weight in kg, and BMI index. The BMI index is 0 to 5, with 0 being extremely weak, 1 being weak, 2 being normal, 3 being overweight, 4 being obese, and 5 being extremely fat. On these criteria, we will do a few hypothesis tests.

The data was first read into R studio. We use the clean_names function from the janitor library to make our column names R compatible. Next, we check the data structure for any anomalies. For the analysis, every sort of data is suitable. The next step is to do a Null check after checking our data for any blanks or nans. The lack of blanks, nulls, or nan indicates that the data is clean, as can be shown.

Figure 1: Data Loading and Cleaning

```
> bmi <- read.csv("bmi.csv")
> p_load(janitor)
> bmi = clean_names(bmi)
> #Checking data structure
> str(bmi)
'data.frame':   500 obs. of  4 variables:
 $ gender: chr  "Male" "Male" "Female" "Female" ...
 $ height: int  174 189 185 195 149 189 147 154 174 169 ...
 $ weight: int  96 87 110 104 61 104 92 111 90 103 ...
 $ index : int  4 2 4 3 3 3 5 5 3 4 ...
> #Checking for Missing data
> sapply(bmi, function(x) sum(x == ""))
gender height weight  index
     0      0      0      0
> sapply(bmi, function(x) sum(x == "nan"))
gender height weight  index
     0      0      0      0
> #Checking for Nan values in data
> colSums(is.na(bmi))
gender height weight  index
     0      0      0      0
```

Subsequently, we apply a filter to the data in order to generate two subsets: one for every female and another for every man. When we applied descriptive statistics to these categories, we found that the mean height of males

Figure 2: Descriptive Statistics

```
> #Subsetting data based on gender
> male_data <- filter(bmi, gender == "Male")
> female_data <- filter(bmi, gender == "Female")
> #Descriptive statistics for each gender
> describe(male_data)
        vars   n   mean    sd median trimmed   mad min max range  skew kurtosis   se
gender*    1 245   1.00  0.00      1    1.00  0.00   1   1     0   NaN      NaN 0.00
height     2 245 169.65 17.07    171  169.87 20.76 140 199    59 -0.13    -1.19 1.09
weight     3 245 106.31 31.83    105  106.64 40.03  50 160   110 -0.05    -1.20 2.03
index      4 245   3.79  1.39      4    3.99  1.48   0   5     5 -0.98    -0.05 0.09
> describe(female_data)
        vars   n   mean    sd median trimmed   mad min max range  skew kurtosis   se
gender*    1 255   1.00  0.00      1    1.00  0.00   1   1     0   NaN      NaN 0.00
height     2 255 170.23 15.71    170  170.40 20.76 140 199    59 -0.09    -1.10 0.98
weight     3 255 105.70 32.96    106  105.86 41.51  50 160   110 -0.03    -1.23 2.06
index      4 255   3.71  1.33      4    3.86  1.48   0   5     5 -0.86    -0.05 0.08
```

measured 169.65 cm, 106.31 kg on average, and 3.79 on average for the BMI. The average height, weight, and BMI of females are 170.23 cm, 105.70 kg, and 3.71 BMI.

We performed three hypothesis tests on the data. Below is a list of each test's outcomes.

Figure 3: Hypothesis testing for BMI index

```
> #Hypothesis testing that the the average BMI index of the sample is less than average male BMI index
> t.test(bmi$index, mu = 3.79, alternative = "less", conf.level = .95)

        One Sample t-test

data:  bmi$index
t = -0.69307, df = 499, p-value = 0.2443
alternative hypothesis: true mean is less than 3.79
95 percent confidence interval:
     -Inf 3.847863
sample estimates:
mean of x
    3.748
```

We are determining if the sample's average BMI index is smaller than 3.79 for our first hypothesis test, and we have selected the 95% confidence interval because it strikes a compromise between certainty and accuracy. The average BMI index of the sample and the average BMI index of men do not significantly vary, according to the null hypothesis (3.79). A more plausible theory is that the average BMI index is below 3.79.

The degree of freedom is 499, the p-value is 0.2443, and the t-value is around -0.69307. Considering the negative t-value and comparatively high p-value, the null hypothesis cannot be rejected with sufficient evidence. There is a large range of conceivable values for the real mean, as indicated by the 95% confidence interval, which spans negative infinity and 3.847863. In conclusion, there is not enough data to reject the null hypothesis that the real mean BMI index is equal to 3.79 based on this one-sample t-test. The observed mean of 3.748 is not statistically different from 3.79, according to the p-value of 0.2443. 3.79 is included in the broad confidence interval, indicating a variety of tenable values for the real mean.

Figure 4: Hypothesis testing for height

```
> #Hypothesis testing that the the average height of the sample is greater than 160
> t.test(bmi$height, mu = 160, alternative = "greater", conf.level = .95)

        One Sample t-test

data:  bmi$height
t = 13.579, df = 499, p-value < 0.00000000000000022
alternative hypothesis: true mean is greater than 160
95 percent confidence interval:
 168.7372      Inf
sample estimates:
mean of x
  169.944
```

We are assessing if the sample's average height is more than 160 in terms of height, and the 95% confidence interval was selected because it strikes a compromise between certainty and precision. The sample's average height is not larger than 160 cm, according to the null hypothesis. An alternative theory is that people are taller on average than 160 cm.

The degree of freedom is 499, the p-value is 0.00000000000000022, and the t-value is around 13.579. There is substantial evidence to reject the null hypothesis due to the big positive t-value and very small p-value. Conclusion further supported by the fact that 160 is not included in the 95% confidence range for the real mean. In conclusion, there is compelling evidence to imply that the sample's real mean height is higher than 160 based on this one-sample t-test. The confidence interval offers a range of reasonable values for the genuine mean that starts above 160, and the sample mean is estimated to be 169.944.

Figure 5: Hypothesis testing for weight

```
> #Hypothesis testing that the the average weight of the sample is equal to average female weight
> t.test(bmi$weight, mu = 105.70, alternative = "two.sided", conf.level = .95)

        One Sample t-test

data:  bmi$weight
t = 0.20715, df = 499, p-value = 0.836
alternative hypothesis: true mean is not equal to 105.7
95 percent confidence interval:
 103.1547 108.8453
sample estimates:
mean of x
      106
```

We are assessing if the sample's average weight is equivalent to the average weight of females, which is 105.70 kg. The 95% confidence interval was selected because it strikes a compromise between certainty and precision. The sample's average weight of 105.70 kg is the null hypothesis in this case. An alternative theory is that 105.70 kg is not the average weight.

The p-value is 0.836, the degree of freedom is 499, and the t-value is around 0.20715. It is not possible to reject the null hypothesis with the t-value being so close to zero and the p-value being so high. There is no evidence to contradict the hypothesised value of 105.7, as the 95% confidence range for the real mean encompasses it. In conclusion, there is not enough data to reject the null hypothesis that the real mean weight is equal to 105.7 based on this one-sample t-test. Given the substantial p-value (0.836), it is unlikely that the observed mean of 106 and 105.7 differ much. 105.7 is included in the confidence interval (103.1547, 108.8453), which shows a range of likely values for the actual mean weight.

Conclusion

After evaluating the BMI data, we were able to examine several statistical outputs of the variables and provide descriptive statistics of the data. To better comprehend t-test results and p-value, we were also able to do hypothesis testing on a few columns. We also discovered the circumstances under which the null hypothesis is accepted or rejected. The null hypothesis should be accepted when the t-value is near to zero or negative and the p-value is reasonably big; the alternative hypothesis should be accepted when the t-value is large and the p-value is extremely tiny.

Citations

R Documentation, An introduction to R. Retrieved 09th June 2024 from https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Related-software-and-documentation

Null hypothesis, Retrieved 09th June 2024 from https://byjus.com/maths/null-hypothesis/