University of Coimbra

Doctoral Program in Information Science and Technology

Statistics

Assignment #3 - Population Testing

Joaquim Pedro Bento Gonçalves Pratas Leitão - 2011150072

January 5, 2017

1 Introduction

The current document is framed in the scope of the third assignment of the Statistics course, taught for the Doctoral Program in Information Science and Technology at the University of Coimbra, during the academic year of 2016/2017.

The current assignment focus on parametric and non-parametric significance tests with respect to the average of a given population of which a sample was observed and collected.

The data collected for this assignment through a survey of fast-food consumption habits in England was used in this study, and the opinions of 344 consumers who reported buying fast food regularly were recorded. The supplied dataset contains the following variables:

- Average number of weekly fast-food purchases in the last month.
- Age of the reporting person, divided in five classes: **1** for people with ages between 15 and 17, **2** for 18-24, **3** for 25-35, **4** for 36-54 and **5** for 55-70.
- Genre of the reporting person.
- Socio-economic level of the reporting person, divided in three classes: **1** for low, **2** for medium and **3** for high.
- Educational qualifications of the reporting person, divided in four classes: **1** for high school level, **2** for technical course, **3** for B.Sc degree and **4** for MSc or PhD.

As already mentioned, the main focus and goal of the current assignment was on the application of adequate parametric and non-parametric significance tests to the collected and observed samples, in order to withdraw conclusions with respect to the entire population under study.

The main variable under analysis is the average number of weekly fast-food purchases, and the three tests proposed for the current assignment relate this variable with some of the remaining collected variables: In the first two tests fast-food purchases are related with the genre of the inquired people and in the final test an analysis of fast-food purchases by age is performed.

As required, all tests were performed with a significance level of 0.05. In addition, all tests documented in this report were performed using the R[1] software environment.

In the remainder of this document the tests proposed in the current assignment will be covered: In section 2 the first two tests are presented and discussed, while section 3 is reserved for the third test.

[1] `https://www.r-project.org/`

2 Fast-Food Purchases by Genre

In the current section the two proposed tests to relate fast-food purchases with the survey respondents' genre are covered.

2.1 Purchases by Women

As a starting point it was intended to analyse the average fast-food purchases made by women, in order to determine whether or not this average value (in the previous month) is significantly higher than a given value, 2.5 in our scenario.

Taking into account that only one variable was being analysed at this point (corresponding to a subset of the collected sample) we are faced with a test for the average of a population, which may or may not be parametric. Therefore, a *Student's T-Test* (T-Test) appears as the most promising test, provided its assumptions can be met.

Since the T-Test assumes the random variable in question follows a normal distribution, the initial step in this problem will be determining whether or not the variable of interest (in this case, the average fast-food purchases made in the last month) follows a normal distribution.

To this end, a *Shapiro-Wilk* test for normality in the variable of interest was performed. A *p-value* of $7.191026e^{-12}$ was registered in this test. As this *p-value* is considerably smaller than the required significance level, the test's null hypothesis (stating that the variable under analysis follows a normal distribution) is rejected, at the significance level of 0.05.

These results are further supported by the histogram of the average fast-food purchases by women, presented in figure 1.

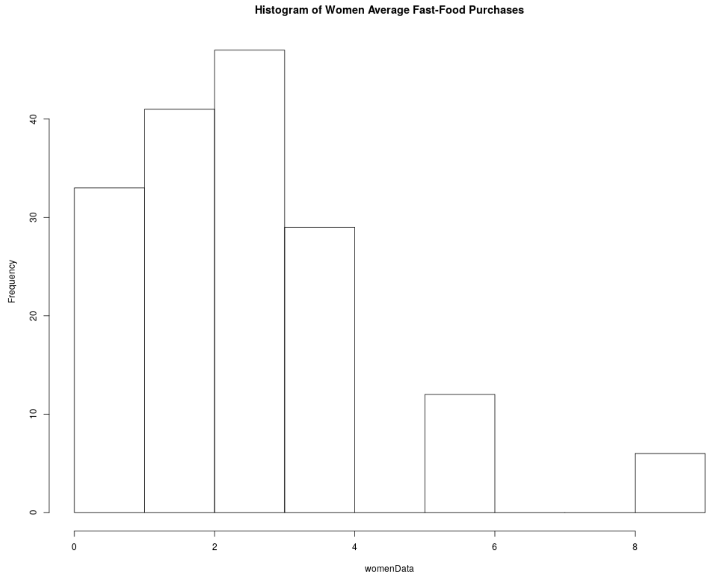

Figure 1: Histogram of fast-food purchases by women.

A practical result of the application of the *Shapiro-Wilk* test for the fast-food purchases by women implies that this variable cannot be accepted to follow a normal distribution.

Even though normality cannot be assured for our study variable, the T-Test can still be applied to this variable, provided its number of observations is higher than 30. In this case we obtain an approximate *p-value* for the test. Considering that a total of 168 women answered the survey, the T-Test can still be applied to the fast-food purchases made by women (as $168 > 30$).

As a result, the following hypothesis were considered in the T-Test:

$$H0: \ m_X = 2.5 \ \ vs \ \ H1: m_X > 2.5$$

where *X="Average number of weekly fast-food purchases made by women in the last month"*.

For this test an approximate *p-value* of 0.2779 was registered. As this value is considerably higher than the required significance level (0.05) the null hypothesis is accepted, at this significance level. Such hypothesis states that the average number of weekly fast-food purchases made by women in the last month is 2.5.

Based on what was presented in this subsection and on the results of the performed hypothesis tests, the main conclusion to be withdrawn at this point is that it cannot be concluded that the average number of weekly fast-food purchases made by women in the last month is significantly higher than 2.5.

2.2 Purchases by Women and Men

Extending the analysis performed in the previous subsection, the second proposed exercise intended to compare the average fast-food purchases by both women and men, in order to determine if the average purchases by women is significantly higher than the ones made by men.

In such an exercise it is intended to compare the means of two populations: The average purchases by women and by men. Therefore, a *Student's T-Test* (T-Test) appears as the most promising significance test to be performed. Similarly to what was done in the previous subsection, prior to the execution of the test, its assumptions must be met.

When comparing the means of two populations the T-Test has two possible formulations, one applied when the collected samples are independent, and another when in the presence of paired samples. As purchases by women and men have no relation between then, therefore there is no pairing between collected samples of the two populations, in the current exercise a T-Test for independent samples is considered.

Besides assuming the independence of the corresponding random variables, this T-Test also assumes the normality of both variables. When one of the variables (or both) does not follow a normal distribution the test can still be applied, provided a high number of samples is collected for both variables (higher than 30). In this scenario, an approximated *p-value* is obtained.

To this end, the analysis in this exercise started by performing a *Shapiro-Wilk* test for normality in both variables (purchases made by women and by men). Since the *Shapiro-Wilk* test was performed in the previous exercise for the average fast-food purchases made by women, in the current exercise this test was only performed for the average fast-food purchases made by men.

The normality test registered a *p-value* of $2.515e^{-10}$, which is considerably smaller than the required significance level. As a result, the test's null hypothesis (stating that the variable under analysis follows a normal distribution) is rejected, at the significance level of 0.05. This result is further supported by the histogram of the average fast-food purchases by men, presented in figure 2.

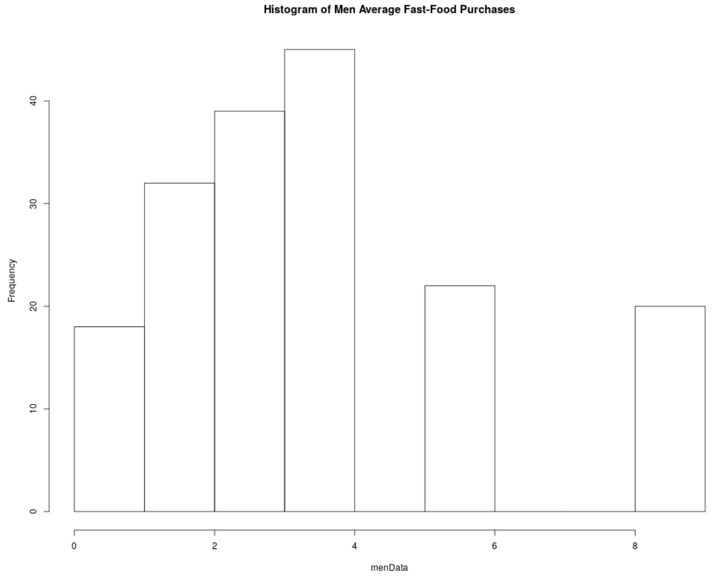

Figure 2: Histogram of fast-food purchases by men.

Since both variables being considered in this exercise registered considerably small *p-values* in their normality tests, at the required significance level, their normality cannot be accepted; however, taking into account that the number of collected samples for each variable is considerably higher (168 for women and 176 for men, which are both higher than 30), the T-Test can still be applied and an approximate *p-value* will be obtained.

As a result, the following hypothesis were considered in the T-Test:

$$H0 : \ m_X - m_Y = 0 \ \ vs \ \ H1 : m_X - m_Y > 0$$

where *X="Average number of fast-food purchases per week made by women in the last month"* and *Y="Average number of fast-food purchases per week made by men* *in the last month".*

The T-Test for independent samples considers two distinct test statistics to be computed, depending on the equality (or not) of the variances of the random variables. Therefore, the next step to be performed in this exercise is to analyse the variance equality between the two considered variables. Such a task is accomplished by performing the *Levene's Test* for variance equality. In this test, a *p-value* of 0.002105 was registered. Since this value is considerably smaller than the required significance level, the null hypothesis is rejected and the variances must be considered different, at the required significance level.

In this scenario (independent samples with different variances) the T-Test with Welch modification is performed. An approximate *p-value* of 0.999994 was registered. As this value is considerably higher than the required significance level, the null hypothesis of this test is accepted (again, at the required significance level).

Based on what was presented in the current subsection, at the significance level of 0.05, one can state that the average purchases made by women and men in the previous month are not significantly different (and, therefore, the purchases made by women are not significantly higher than the ones made by men).

3 Fast-Food Purchases by Age

In the final exercise average fast-food purchases are intended to be analysed according to the consumers' age group, in order to determine if a significant difference is registered between the five age groups considered.

When comparing two or more groups with respect to the location, *ANOVA* appears as the parametric test of choice, with *Kruskal-Wallis* being its non-parametric equivalent.

In order to be applied, ANOVA assumes the normality of each group being considered and the homogeneity of their variances (that is, the variances of the groups must be equal). Similarly to the approach carried out in previous exercises, the first step to compare the average fast-food purchases by consumers' age group is to determine whether or not each group follows a normal distribution.

To accomplish this, observed data for each group is collected and the *Shapiro-Wilk* normality test is performed on each of these groups. In the current work, the normality test was first applied to age group number **2** (for people with ages ranging from 18 to 24 years old), registering a *p-value* of 0.0003072. As this is considerably smaller than the required significance level, the fast-food purchases for people in the second group cannot be accepted to follow a normal distribution, at the significance level of 0.05.

As a result, the first assumption of the ANOVA cannot be met (the normality of each group) and therefore this parametric test is not applicable to the collected data.

Its non-parametric alternative, the *Kruskal-Wallis* is thus applied to our collected data.

When applying the *Kruskal-Wallis* test for mean equality, a *p-value* of $4.751e^{-}06$ was registered. As this is considerably smaller than the required significance level the null hypothesis stating that all means are equal is rejected. Therefore, at the significance level of 0.05, fast-food purchases differ according to the consumers' age group.