University of Coimbra

Doctoral Program in Information Science and Technology

Statistics

Assignment #3 - Population Testing

Joaquim Pedro Bento Gonçalves Pratas Leitão - 2011150072

January 16, 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.

2.1.1 Approach to the Problem

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 *Kolmogorov-Smirnov* test with *Lilliefors correction* for normality in the variable of interest was performed. Considering that a sample larger than 30 was collected ($168 > 30$) this normality test was select. If a sample with at most 30 elements had been collected, a *Shapiro-Wilk* test for normality would have been applied instead.

A *p-value* of $2.970296e^{-26}$ was registered in the performed 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.
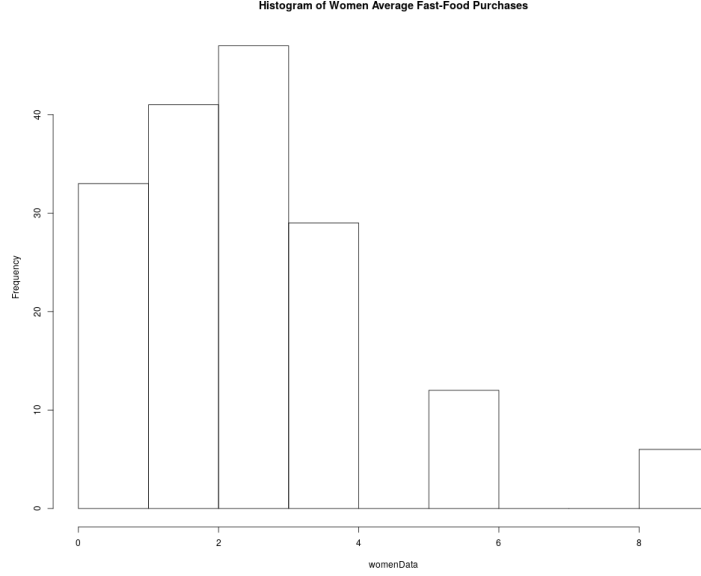
Figure 1: Histogram of fast-food purchases by women.

A practical result of the application of the *Kolmogorov-Smirnov* test with *Lilliefors correction* 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.

2.1.2 R Implementation

In listing 1 the R implementation of the detailed approach is presented. This listing contains only the core commands developed in the course of this work; as such, additional commands can be found in the R scripts provided with this report (namely commands used for debugging or for presenting the obtained results to the user).

Listing 1: AnaylseWomenFastFood.R

```R
AnaylseWomenFastFood <- function(dataset) {
  # Analyses the average number of weekly purchases made by
  # women in the last month and determines if its average
  # level is significantly higher than 2.5 .
  #
  # Args:
  #    dataset: A dataframe with the observed data.
  #

  significanceLevel = 0.05
  confidenceLevel = 1 - significanceLevel

  # Filter data: Only collect information for the women and
  # then select only the "Compras" field
  womenData <- subset(dataset, Sexo=='2')
  womenData <- womenData$Compras

  # Just plot the histogram of the desired data
  png('images/Histogram_WomenData.png', width=1000,
      height=800)
  hist(womenData,
       main='Histogram of Women Average Fast-Food Purchases')

  # Test for normality with Kolmogorov-Smirnov Test with
  # Lilliefors Correction
  lillieResult <- lillie.test(womenData)

  # Compare the p-value (lillieResult$p.value) with the
  # significance level.

  # The obtained p-value for the K-S Test with Lilliefors
  # Correction is very small (7.191e-12) therefore, at
  # the significance level of 0.05, we reject the null
  # hypothesis stating that the observed data follows a
  # normal distribution.
  # Nevertheless, since the length of the observed data is
  # considerably higher than 30 (168 > 30) we can still
  # apply the T-Test for the mean of a population and
  # obtain an approximate p-value.
```

```
  # In the T−Test for the mean of a population we are
  # considering the following hypothesis:
  #     H0: m = 2.5
  #     H1: m > 2.5
  tResult <− t.test(womenData, alternative='greater',
                  mu=2.5, paired=FALSE,
                  conf.level=confidenceLevel)

  # Compare the p−value (tResult$p.value) with the
  # significance level.

  # We obtain an approximate p−value of 0.2779 for the
  # T−Test. For a significance level of alfa = 0.05 we
  # verify that p−value > alfa. Therefore, and at the
  # referred significance level, we accept the null
  # hypothesis stating that m = 2.5

  dev.off()
}
```

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.

2.2.1   Approach to the Problem

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 assump-
tions must be met.

When comparing the means of two populations the T-Test has two possible formu-
lations, 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 *Kolmogorov-Smirnov* test for normality with *Lilliefors correction* in both variables (purchases made by women and by men). Since this 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 $2.071883e^{-18}$, 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.
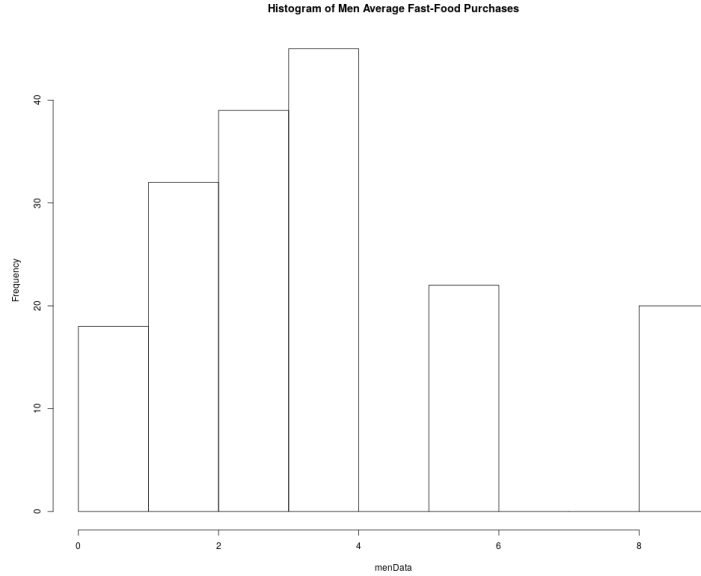

Figure 2: Histogram of fast-food purchases by men.

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:

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

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).

2.2.2 R Implementation

In listing 2 the *R* implementation of the detailed approach is presented. Similarly to the previous problem, this listing contains only the core commands developed in the course of this work. Additional commands can be found in the *R* scripts provided with this report.

Listing 2: AnaylseWomenMenFastFood.R

```
AnalyseWomenMenFastFood <- function ( dataset ) {
  # Analyses the average number of weekly purchases made by
  # women and men in the last month and determines if the
  # average level of women purchases is significantly higher
  # than that of the man.
  #
  # Args:
  #   dataset: A dataframe with the observed data.

  library ( car )

  significanceLevel = 0.05
  confidenceLevel = 1 - significanceLevel

  # X: "Average number of fast-food purchases per week made
  #     by women in the last month"
  # Y: "Average number of fast-food purchases per week made
  #     by men in the last month"
```

```r
#
# We want to test H0: mx - my = 0 vs H1: mx - my > 0

# Filter Women data
womenData <- subset(dataset, Sexo=='2')
womenData <- womenData$Compras

# Filter Men data
menData <- subset(dataset, Sexo=='1')
menData <- menData$Compras

# Just plot the histogram of the desired data
png('images/Histogram_MenData.png', width=1000, height=800)
hist(menData,
     main='Histogram of Men Average Fast-Food Purchases')

# Test X and Y for normality
# In the previous exercise we have already tested X for
# normality and, according to the K-S Test with Lilliefors
# Correction at a significance level of 0.05 we reject the
# null hypothesis stating that the data follows a normal
# distribution.
# Therefore, in this exercise, we just need to perform
# the same test for Y.
lillieResult <- lillie.test(menData)

# Compare the p-value (lillieResult$p.value) with the
# significance level.

# The K-S Test with Lilliefors Correction for normality
# for the Y variable returns a p-value of 2.071883e-18.
# Indeed, this is a very small value which, at the
# significance level of 0.05, makes us reject the null
# hyphtoesis stating that the data follows a normal
# distribution.
#
# Even though both X and Y variables do not follow a normal
# distribution, since the length of the observed data for
# both variables is considerably higher than 30 (168 for
# women and 176 for men) we can still aplly the T-Test
# for the mean difference of two populations, obtaining
# an approximate p-value.
#
# Furthermore, because X and Y are independent, we will be
# applying the T-Test for the mean difference of two
# independent populations. The next step in this exercise
```

```r
    # is to evaluate whether or not the variances of X and Y
    # are equal. Such evaluation can be achieved by performing
    # the Levene's Test.

    # Perform the Levene's Test for Variance Equality
    temp <- data.frame(Compras=dataset$Compras,
                       Sexo=factor(dataset$Sexo))
    leveneResult <-leveneTest(Compras ~ Sexo, data=temp)

    # Get p-value from Levene Test
    leveneP <- (leveneResult$'Pr(>F)')[1]

    if (leveneP <= significanceLevel) {
      cat('\n     Result: Variances of X and Y are different\n')
      varianceEquality = FALSE
    } else {
      cat('\n     Result: Variances of X and Y are equal\n')
      varianceEquality = TRUE
    }

    # Perform T-Test according to the Levene's Test results
    #     H0: mx - my = 0
    #     H1: mx - my > 0
    tResult <- t.test(womenData, menData,
                      alternative='greater',
                      mu=0, paired=FALSE,
                      var.equal=varianceEquality,
                      conf.level=0.95)


    # Compare the p-value (tResult$p.value) with the
    # significance level.

    dev.off()
}
```

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.

3.0.1 Approach to the Problem

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 or *Kolmogorov-Smirnov* test for normality with *Lillierfors correction* were performed on each of these groups. The choice between the tests to apply was made based on the number of observed samples in each group: For groups where at most 30 samples were observed, the *Shapiro-Wilk* test was applied; in groups where more than 30 samples were observed, the *Kolmogorov-Smirnov* test with *Lilliefors correction* was applied.

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 $1.26937327425704e^{-7}$. 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 0.05^2.

As a result, the first assumption of the ANOVA cannot be met (the normality of each group). Even when the normality of the groups cannot be verified, the ANOVA has been proven to be resistant to deviations from the normality of the groups. As a result, it might be worth verifying additional requirements considered in this test, if the cardinality of each group is high enough. Unfortunately, group 1 has a very low cardinality (only 18) elements, 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 $4.751e^{-6}$ 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.

3.0.2 R Implementation

In listings 3 and 4 the *R* implementation of the detailed approach is presented. Similarly to the previous problem, these listing contain only the core commands

---

[2]Even when performing the adequate normality tests to all the considered groups, in no test the null hypothesis considering the normality of the data was accepted.

developed in the course of this work. Additional commands can be found in the R scripts provided with this report.

Listing 3: AnalyseAgeGroupFastFood.R

```r
AnalyseAgeGroupFastFood <- function(dataset) {
  # Analyses the average weekly number of purchases in the
  # last month and determines if it is significantly
  # different according to the age group of the people that
  # answered the survey.
  #
  # Args:
  #   dataset: A dataframe with the observed data.
  #

  # 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.
  # Therefore, in this approach, we will start by checking
  # whether or not the ANOVA can be applied. If its
  # assumptions cannot be verified then we will move on to
  # the Kruskal-Wallis test.

  # Start by getting all the age groups
  ageGroups <- unique(dataset$Idade)

  significanceLevel = 0.05

  # Check if we can perform ANOVA!
  canPerformAnova <- TestIfAnova(dataset, ageGroups,
                                 significanceLevel)

  # Regardless of the type of test that we perform, we
  # will consider the following two hypothesis:
  # H0: m1 = m2 = ... = mr, for r the number of groups
  # H1: The means of the groups are not all equal
  #

  if (canPerformAnova) {
    # ANOVA

    testData <- c()
    testGroup <- c()

    for (i in ageGroups) {
      groupData <- subset(dataset, Idade==i)
      groupData <- groupData$Compras
```

```r
      testData <- c(testData, groupData)
      testGroup <- c(testGroup, rep(i, length(groupData)))
    }

    dataF <- data.frame(y=testData, group=factor(testGroup))
    fitAnova <- lm(formula=testData ~ testGroup, data=dataF)

    testResult <- anova(fitAnova)

    pValue <- testResult$`Pr(>F)`
    pValue <- pValue[1]

  } else {
    # Kruskal-Wallis

    testData <- list()

    for (i in ageGroups) {
      groupData <- subset(dataset, Idade==i)
      groupData <- groupData$Compras

      testData[[i]] <- groupData
    }

    testResult <- kruskal.test(testData)

    pValue <- testResult$p.value
  }

  # Compare the p-value (tResult$p.value) with the
  # significance level.
}
```

```r
TestIfAnova <- function(dataset, groups, significanceLevel) {
  # Given a dataset, the current function checks if all the
  # required assumptions to perform an Analysis of Variance
  # (ANOVA) are met or not.
  #
  # Args:
  #   dataset: A dataframe with the observed data.
  #   groups: The different groups considered in this
  #           exercise.
  #   significanceLevel: The significance level to
  #                      consider in the tests.
  #
```

```r
# Returns:
#   A boolean value, indicating if the ANOVA can, or
#   cannot, be performed.

library(car)
library(nortest)

# Let numberAgeGroups <- length(groups).
# We will create "numberAgeGroups" groups. These will be
# our Xi. Each group will have a dimension ni (for i
# ranging from 1 to "numberAgeGroups").
# For each group we have observed a sample
# (xi_1, ... , xi_ni), and each xi_j is an observation of
# a random variable Xi_j.
# For each of these variables the following properties will
# have to be verified:
#   1) Each random variable Xi_j follows a Normal
#      Distribution.
#   2) All "numberAgeGroups" samples are independent from
#      each other.
#   3) All "numberAgeGroups" samples have the same
#      variance.
#

stop <- FALSE

# Property 1: Each random variable Xi_j follows a normal
# distribution.
# To verify this property we must check, for each group,
# is the observed sample (xi_1, ... , xi_ni) can be
# obtained from a Normal Distribution.
for (i in groups) {
  # Get data in the current group
  groupData <- subset(dataset, Idade==i)
  groupData <- groupData$Compras

  # For the current group test the normality with the
  # K-S Test with Lilliefors Correction or the
  # Shapiro-Wilk Test for normality
  if (length(groupData) > 30) {
    cat(paste('Group ', i, ' has length higher than 30 (ni=',
              length(groupData), ')\n', sep=''))
    testResult <- lillie.test(groupData)
  } else {
    cat(paste('Group ', i, ' has length less or equal to 30 (ni=',
              length(groupData), ')\n', sep=''))
```

```r
      testResult <- shapiro.test(groupData)
    }

    # Inspect the obtained p-value
    if (testResult$p.value <= significanceLevel) {
      cat(paste('Group ', i, ' does not follow a Normal ',
                'Distribution as a p-value of ',
                testResult$p.value,
                ' was registered\n', sep=''))
      stop <- TRUE
    } else {
      cat(paste('Group ', i, ' follows a normal ',
                'distribution\n', sep=''))
    }
  }

  if (stop) {
    return(FALSE)
  }

  # Property 2: Yes, the samples are independent from each
  # other

  # Property 3: Test if all the "numberAgeGroups" samples
  # have the same variance
  if (length(groups) > 1) {
    # Perform Levene Test for homogeneity of variance
    # accross the groups
    temp <- data.frame(Compras=dataset$Compras,
                       Idade=factor(dataset$Idade))
    leveneResult <- leveneTest(Compras ~ Idade, data=temp)

    # Get p-value from Levene Test
    leveneP <- (leveneResult$`Pr(>F)`)[1]

    if (leveneP <= significanceLevel) {
      cat('The variances of the groups are different')
      return(FALSE)
    } else {
      cat('The variances of the groups are the same')
      return(TRUE)
    }
  }

  return(TRUE)
}
```