R Assignment - 3

Jasmeet Singh Saini - 0758054

2023-03-16

Question 1 - Pizza McPizza Face

A frozen pizza maker produces 5 types of pizza: cheese, pepperoni, vegetarian, deluxe, and meat lovers. A random sample of transactions containing frozen pizza was examined from a local grocery store and the numbers of each pizza type recorded.

a) The pizza maker wants to know if the proportions of pizza sold matched the preferences found in a large national poll. Using the information in Table 1, answer this question using a significance level of $\alpha = 0.05$. What test should you use?

Table 1: Pizza sales (sold column) along with preferences from a national poll (poll column).

Pizza Type	Sold	Poll
Cheese	24	0.15
Pepperoni	65	0.40
Vegetarian	12	0.06
Deluxe	30	0.15
Meat	39	0.24

Solution:

Firstly, let's construct the table:

```
table_one = data.frame( pizza_type = c('Cheese', 'Pepperoni', 'Vegetarian', 'Deluxe', 'Meat'),
                   sold = c(24,65,12,30,39), poll = c(0.15,0.40,0.06,0.15,0.24))
table_one
```

```
##   pizza_type sold poll
## 1     Cheese   24 0.15
## 2  Pepperoni   65 0.40
## 3 Vegetarian   12 0.06
## 4     Deluxe   30 0.15
## 5       Meat   39 0.24
```

The objective of this question is to determine whether the distribution of pizza sales in a store matches with the claimed distribution of pizza sales reported in a large nation poll. To achieve this, we will use the χ^2 **goodness-of-fit test**, which tells, whether or not the observed frequency distribution fits a particular claimed one.

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

$$H_0 : \text{The Proportion of pizzas sold matches with the national poll.}$$

vs.

$$H_A : \text{The Proportion of pizzas sold does not matches with the national poll.}$$

Assumptions

To apply CLT, the below conditions must be met:

a. **Independence** : Each observation is randomly sampled independent of the others. Therefore Independence condition is met.

b. **Success / Failure Conditions** : In order to check this condition, we need to each category must have at least 5 expected cases.

Let's check the conditions:

```
testing_a = chisq.test(table_one$sold, p = table_one$poll, correct = TRUE)
testing_a$expected
```

```
## [1] 25.5 68.0 10.2 25.5 40.8
```

```
all(testing_a$expected >= 5)
```

```
## [1] TRUE
```

Since, all values are greater than or equal to 5. Thus, the condition is satisfied. Now, let's analyze the result of chi-squared test to make our decision.

Step 2 : Test Statistic and p-value

```
str(testing_a)
```

```
## List of 9
##  $ statistic: Named num 1.41
##   ..- attr(*, "names")= chr "X-squared"
##  $ parameter: Named num 4
##   ..- attr(*, "names")= chr "df"
##  $ p.value  : num 0.842
##  $ method   : chr "Chi-squared test for given probabilities"
##  $ data.name: chr "table_one$sold"
##  $ observed : num [1:5] 24 65 12 30 39
##  $ expected : num [1:5] 25.5 68 10.2 25.5 40.8
##  $ residuals: num [1:5] -0.297 -0.364 0.564 0.891 -0.282
##  $ stdres   : num [1:5] -0.322 -0.47 0.581 0.967 -0.323
##  - attr(*, "class")= chr "htest"
```

Step 3 : Statistical Decision

Hence, the test statistic(χ^2) is **1.41** and $p - value(0.84) > 0.05$, we will fail to reject the null hypothesis.

Step 4 : Conclusion

From our sample, there is enough evidence to support the null hypothesis and therefore, we would conclude that the claim, "the proportion of pizzas sold **matches** with the national poll (poll column)", is accurate.

b) The pizza maker let the internet choose their next three types of pizza. They should have known better1 and the internet decided on: cotton candy, BBQ chicken, and taco. Randomly sampled transactions were selected from a local grocery store and sales of these new flavours were compared with internal marketing data for the expected sales at the same store during a one week period.

Use the information in Table 2 along with a goodness-of-fit test to decide if the internal marketing data was correct ($\alpha = 0.1$).

Table 2: Pizza sales (sold column) and expected sales (expected column) for the three new types.

Pizza Type	Sold	Expected
Cotton Candy	0	2
BBQ Chicken	12	10
Taco	5	3

Solution:

Firstly, let's construct the table:

```
table_two = data.frame(pizza = c('Cotton Candy', 'BBQ Chicken', 'Taco'),
                       sold = c(0,10,5),
                       expected_sales = c(2,10,3))
table_two
```

```
##          pizza sold expected_sales
## 1 Cotton Candy    0              2
## 2  BBQ Chicken   10             10
## 3         Taco    5              3
```

We will use the χ^2 **goodness-of-fit test**.

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : The Proportion of pizzas sold matches with their expected sales.

vs.

H_A : The Proportion of pizzas sold does not matches with their expected sales.

Assumptions

To apply CLT, the below conditions must be met:

a. **Independence** : Each observation is randomly sampled independent of the others. Therefore Independence condition is met.

b. **Success / Failure Conditions** : In order to check this condition, we need to each category must have at least 5 expected cases.

Let's check the condition:

```r
testing_b = chisq.test(table_two$sold,
                       p = prop.table(table_two$expected_sales),
                       correct = TRUE)
```

```
## Warning in chisq.test(table_two$sold, p = prop.table(table_two$expected_sales),
## : Chi-squared approximation may be incorrect
```

```r
testing_b$expected
```

```
## [1]  2 10  3
```

```r
all(testing_b$expected >=5)
```

```
## [1] FALSE
```

Since, all the values are **not** greater than or equal to 5. Thus, the condition is not satisfied.

Now, we will use simulations non-parametric approach for the test now.

Step 2 : Test Statistic and p-value

Now, we need to perform p-value simulations for chi-squared test to make our decision.

```r
testing_b = chisq.test(table_two$sold,
                       p = prop.table(table_two$expected_sales),
                       correct = TRUE, simulate.p.value = TRUE)
str(testing_b)
```

```
## List of 9
##  $ statistic: Named num 3.33
##   ..- attr(*, "names")= chr "X-squared"
##  $ parameter: Named logi NA
##   ..- attr(*, "names")= chr "df"
##  $ p.value  : num 0.181
##  $ method   : chr "Chi-squared test for given probabilities with simulated p-value\n\t (based on 200
##  $ data.name: chr "table_two$sold"
##  $ observed : num [1:3] 0 10 5
##  $ expected : num [1:3] 2 10 3
##  $ residuals: num [1:3] -1.41 0 1.15
##  $ stdres   : num [1:3] -1.52 0 1.29
##  - attr(*, "class")= chr "htest"
```

Step 3 : Statistical Decision

Hence, the test statistic(χ^2) is **3.33** and $p - value(0.17) > 0.05$, we would **accept** the null hypothesis.

Step 4 : Conclusion

From our sample, there is enough evidence to support the null hypothesis and therefore, we would conclude that the claim, "The Proportion of pizzas sold matches with their expected sales", is accurate.

Question 2 - Cats and Dogs and Fish (oh my!)

You want, nay need, to answer the question that plagues every generation: does a voter's party affiliation affect their pet preference? You conduct a simple random sample of the Canadian electorate and ask participants, "What is your favourite animal?" and "What is your party affiliation?". The information you collect is summarized as:

- **Dog preferred and Liberal: 20**
- **Dog preferred and Conservative: 10**
- **Dog preferred and NDP: 25**
- **Cat preferred and Liberal: 15**
- **Cat preferred and Conservative: 23**
- **Cat preferred and NDP: 20**
- **Fish preferred and Liberal: 10**
- **Fish preferred and Conservative: 30**
- **Fish preferred and NDP: 10**

a) Present this information in a nice table. A table of contingencies perhaps; one might say a contingency table.

Solution:

Firstly, we should construct the contingency table for our data.

```
animal  =  c('Dog', 'Cat', 'Fish')
party   = c('Liberal', 'NDP', 'Conservative')
# Creating the contingency table
contingency_table = table(animal, party)

# Inserting the values into the table
contingency_table['Dog', 'Liberal'] = 20
contingency_table['Dog', 'Conservative'] = 10
contingency_table['Dog', 'NDP'] = 25
contingency_table['Cat', 'Liberal'] = 15
contingency_table['Cat', 'Conservative'] = 23
contingency_table['Cat', 'NDP'] = 20
contingency_table['Fish', 'Liberal'] = 10
contingency_table['Fish', 'Conservative'] = 30
contingency_table['Fish', 'NDP'] = 10
# view the contingency table
contingency_table
```

```
##       party
## animal Conservative Liberal NDP
##   Cat            23      15  20
##   Dog            10      20  25
##   Fish           30      10  10
```

b) What type of test should be used t answer this question?

Solution:

We need to determine whether a party's affiliation affects the pet preference. Thus, a **chi-square test of Independence** would be an appropriate strategy.

c) Test the question of whether party affiliation affect the probability of type of pet. Since this is such an important question, use the significance level $\alpha = 0.01$.

Solution:

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

$$H_0 : \text{The Party affiliation does not affect the probability of type of pet.}$$
$$\text{vs.}$$
$$H_A : \text{The Party affiliation affect the probability of type of pet.}$$

Assumptions

To apply CLT, the below conditions must be met:

a. **Independence** : Each observation is randomly sampled independent of the others. Therefore Independence condition is met.

b. **Success / Failure Conditions** : In order to check this condition, we need to each category must have at least 5 expected cases.

Let's check the condition:

```
testing_two_c = chisq.test(contingency_table, correct = TRUE)
testing_two_c$expected
```

```
##       party
## animal Conservative  Liberal      NDP
##   Cat      22.41718 16.01227 19.57055
##   Dog      21.25767 15.18405 18.55828
##   Fish     19.32515 13.80368 16.87117
```

```
all(testing_two_c$expected >=5)
```

```
## [1] TRUE
```

As, all values are greater than or equal to 5, the condition is satisfied.

Step 2 : Test Statistic and p-value

Now, we need to analyze the result of chi-squared test to make our decision.

```
str(testing_two_c)
```

```
## List of 9
##  $ statistic: Named num 19.6
##   ..- attr(*, "names")= chr "X-squared"
##  $ parameter: Named int 4
##   ..- attr(*, "names")= chr "df"
##  $ p.value  : num 0.000611
##  $ method   : chr "Pearson's Chi-squared test"
##  $ data.name: chr "contingency_table"
##  $ observed : 'table' num [1:3, 1:3] 23 10 30 15 20 10 20 25 10
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ animal: chr [1:3] "Cat" "Dog" "Fish"
##   .. ..$ party : chr [1:3] "Conservative" "Liberal" "NDP"
##  $ expected : num [1:3, 1:3] 22.4 21.3 19.3 16 15.2 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ animal: chr [1:3] "Cat" "Dog" "Fish"
##   .. ..$ party : chr [1:3] "Conservative" "Liberal" "NDP"
##  $ residuals: 'table' num [1:3, 1:3] 0.123 -2.442 2.428 -0.253 1.236 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ animal: chr [1:3] "Cat" "Dog" "Fish"
##   .. ..$ party : chr [1:3] "Conservative" "Liberal" "NDP"
##  $ stdres   : 'table' num [1:3, 1:3] 0.196 -3.83 3.723 -0.37 1.785 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ animal: chr [1:3] "Cat" "Dog" "Fish"
##   .. ..$ party : chr [1:3] "Conservative" "Liberal" "NDP"
##  - attr(*, "class")= chr "htest"
```

Step 3 : Statistical Decision

Hence, the test statistic(χ^2) is **19.6** and $p - value(0.0006) < 0.01$, we would **reject** the null hypothesis.

Step 4 : Conclusion

From our sample, there is not enough evidence to support the null hypothesis and therefore, we would conclude that the claim, "The Party affiliation affect the probability of type of pet", is accurate.

Question 3 - Memories

An experiment was conducted to test whether ginkgo biloba would improve memory. 203 participants were randomly assigned to take ginkgo biloba supplements or a placebo. Memory was tested using standard methods and the changes in scores were recorded as continuous numeric quantities. These are available in memory.csv.

a) Load the dataset contained in memory.csv into R using read.csv().

Solution:

Let's load the dataset using read.csv() into environment variable.

```
# Importing the dataset
memory_data <- read.csv("memory.csv")
```

b) State the means and standard deviations for both groups.

Solution:

The means and standard deviations for both groups, that is, **Gingko Biloba** and **Placebo** are given below:

```
# Mean of Gingko Biloba
round(mean(memory_data$Gingko, na.rm = TRUE),4)
```

```
## [1] 4.2308
```

```
# Standard Deviation of Gingko Biloba
round(sd(memory_data$Gingko, na.rm = TRUE),4)
```

```
## [1] 5.0324
```

The *mean* and *standard deviation* for **Gingko Biloba** is **4.2308** and **5.0324** respectively.

```
# Mean of Placebo
round(mean(memory_data$Placebo, na.rm = TRUE),4)
```

```
## [1] 5.2222
```

```
# Standard Deviation of Placebo
round(sd(memory_data$Placebo, na.rm = TRUE),4)
```

```
## [1] 4.1171
```

The *mean* and *standard deviation* for **Placebo** is **5.2222** and **4.1171** respectively.

c) Which version of two sample test for means is appropriate to determine whether ginkgo biloba improves memory scores? (We know of three versions: paired; independent equal variance; independent unequal variance)

Solution:

For Two Sample, **Independent Equal Variance t-test** for means is appropriate to determine whether ginkgo biloba improves memory scores. Here, we will comparing means of two independent groups, gingko and placebo.

d) Conduct a hypothesis test to see if the mean memory score in the Gingko group is higher than the mean memory score in the Placebo group. Use $\alpha = 0.1$. (This means hypotheses, assumptions, test statistic and *p*-value, statistical decision, conclusion.)

Solution:

Firstly, let's analyze the data:

```
tail(memory_data)
```

```
##     Gingko Placebo
## 99      -6       0
## 100      0      NA
## 101      8      NA
## 102      8      NA
## 103      9      NA
## 104      4      NA
```

The memory data has *NA* values in Placebo. Now, let's remove *NA* values:

```
# Checking Gingko
which(is.na(memory_data$Gingko))
```

```
## integer(0)
```

```
gingko_memory_data <- memory_data$Gingko
```

Here, there are no *NA* values in Gingko group of memory data.

```
# Checking Placebo
placebo_na <- which(is.na(memory_data$Placebo))
placebo_memory_data <- memory_data$Placebo[-placebo_na]
```

Here, we have removed the *NA* values in placebo group.

```
# Checking the length of each group
length(gingko_memory_data)
```

```
## [1] 104
```

```
length(placebo_memory_data)
```

```
## [1] 99
```

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : The mean memory score in Ginkgo is not significantly different from the mean memory score in Placebo.

vs.

H_A : The mean memory score in Ginkgo is significantly higher than the mean memory score in Placebo.

Also, we can say that,

$H_0 : \mu_{ginkgo} \leq \mu_{placebo}$

vs

$H_A : \mu_{ginkgo} > \mu_{placebo}$

we will use a significance level (α) of 0.1.

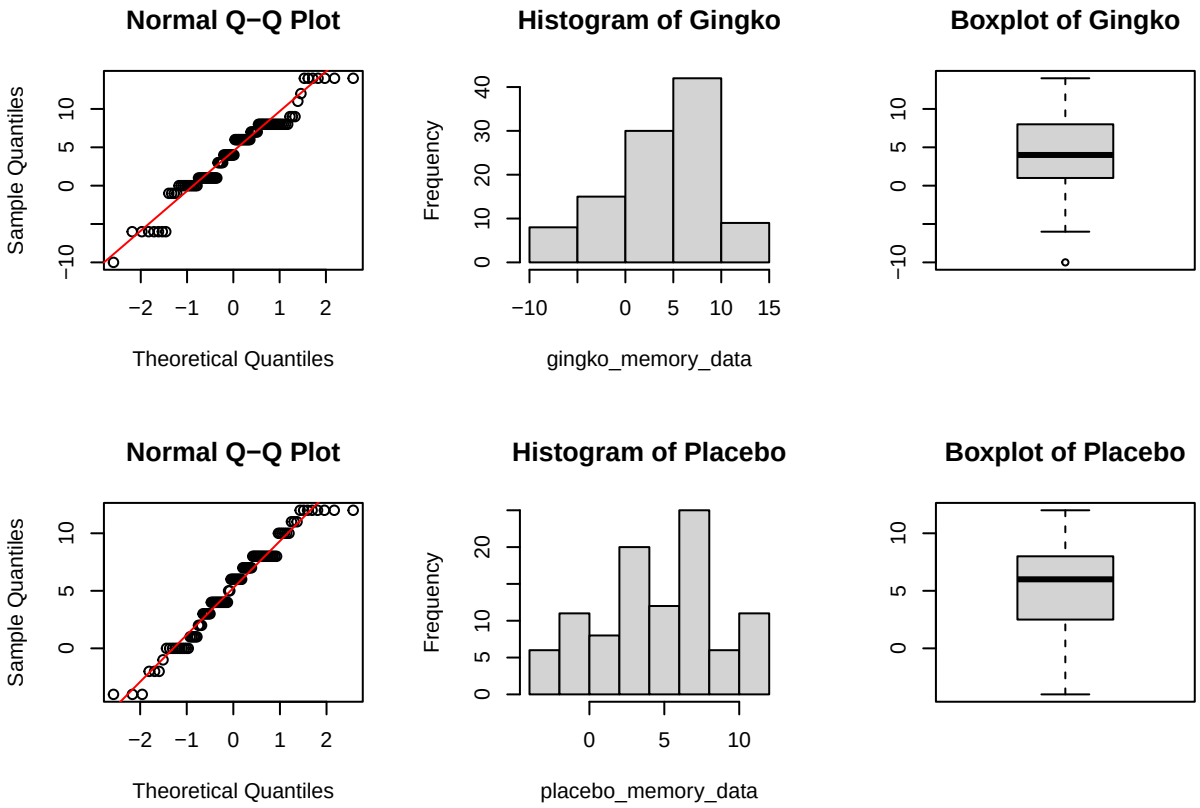
Assumptions

a. **Independence** : The data is randomly sampled and is less than 10% of the population size. Therefore, Independence condition is met.

b. **Normality** : To check whether the data is normally distributed, few visualizations.

```
par(mfrow=c(2,3))
# To visualization and check normality of Gingko group
qqnorm(gingko_memory_data)
qqline(gingko_memory_data, col='red')
hist(gingko_memory_data, main = "Histogram of Gingko")
boxplot(gingko_memory_data, main = "Boxplot of Gingko")

# par(mfrow=c(1,3))
# To visualization and check normality of Placebo group
qqnorm(placebo_memory_data)
qqline(placebo_memory_data, col='red')
hist(placebo_memory_data, main = "Histogram of Placebo")
boxplot(placebo_memory_data, main = "Boxplot of Placebo")
```



Hence, the qqnorm, histogram and boxplot depicts that both the groups, that is, Gingko and Placebo are approximately normal. Thus, the condition is assumptions to be met.

Step 2 : Test Statistic and p-value

```
t_test <- t.test(gingko_memory_data, placebo_memory_data
                 , alternative = "greater"
                 , var.equal = TRUE
                 , conf.level = 0.9)
t_test
```

```
##
##  Two Sample t-test
##
## data:  gingko_memory_data and placebo_memory_data
## t = -1.532, df = 201, p-value = 0.9365
## alternative hypothesis: true difference in means is greater than 0
## 90 percent confidence interval:
##  -1.82355      Inf
## sample estimates:
## mean of x mean of y
##  4.230769  5.222222
```

Hence, the test statistic is **-1.532** and p-value is **0.9365**.

Step 3 : Statistical Decision

As, we know if the p-value is less than or equal to the significance level (α), we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

Here, the $p - value > 0.1$ (level of significance), therefore, we **Accept** the *Null Hypothesis*, H_0.

Step 4 : Conclusion

The t-test results depicts that, the test statistic is t = -1.53 and the p-value is 0.94. Since, the p-value is less than the significance level of 0.1. From our sample, there is enough evidence to support the null hypothesis and therefore, we would conclude that the claim, "The mean memory score in Ginkgo is not significantly different from the mean memory score in Placebo", is accurate.

e) Examine your test statistic value. Based on this and the alternative hypothesis, did you need to run a full statistical analysis to arrive at your conclusion? Describe your rationale briefly.

Solution:

Yes, it is necessary to run a full statistical analysis to determine whether the observed difference in mean memory score between the Ginkgo and Placebo groups is statistically significant.

To arrive at a conclusion about whether this difference is statistically significant, we do need to run a full statistical analysis using the independent samples t-test assuming equal variances and calculate the p-value. The p-value tells us the probability of observing a difference as large as the one we observed or larger, assuming the null hypothesis is true.