Abdulaziz Alqumayzi G200007615 Module-11 CT-4

Rudra S Bandhu

11/27/2021

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(readxl)
```

Import datasets

```
problem_1 <- read_excel("CT4_problems.xlsx", sheet = "problem 1")
problem_2 <- read_excel("CT4_problems.xlsx", sheet = "problem 2", col_types =
c("text","text", "numeric"))
problem_2.1 <- read_excel("CT4_problems.xlsx",sheet= "problem 2.1",
col_types= c("text","text", "numeric"))
problem_3 <- read_excel("CT4_problems.xlsx", sheet = "problem 3")
```

Problem 1

Point 1: If the death day does not depend on the birthday, then it would seem that each of the individuals would be equally likely to fall in any of the 12 categories. Thus, the null hypothesis equal to 1/12

```
model_1 <- chisq.test(problem_1$`Number of Deaths`)
model_1

##
##  Chi-squared test for given probabilities
##
## data:  problem_1$`Number of Deaths`
## X-squared = 17.192, df = 11, p-value = 0.1023
```

Point 2:

Results: X-squared = 17.192, df = 11, p-value = 0.1023

Point 3: p_value is 0.1023

Point 4: The results of this test show the hypothesis that an approaching birthday has no effect on an individual's remaining lifetime. For whereas the data are not quite strong enough (at least, at the 10 percent level of significance) to reject this hypothesis, they are certainly suggestive of its possible falsity.

Problem 2

Part one:

```
model_2<- aov(problem_2$Mileage ~ problem_2$Gasoline + problem_2$Additive)
anova(model_2)
```

```
## Analysis of Variance Table
##
## Response: problem_2$Mileage
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## problem_2$Gasoline   2  1.807  0.9033  0.3798 0.70627
## problem_2$Additive   2 54.860 27.4300 11.5333 0.02184 *
## Residuals            4  9.513  2.3783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Point a: The null hypothesis: gasoline used does not affect the mileage.

the alternative hypothesis: gasoline used does affect the mileage

F-value= 0.3798 p-value= 0.70627

Since p-value > 0.05 at the level of significance(0.05), we fail to reject Ho Conclusion: gasoline used does not affect the mileage

Point b: The null hypothesis: additives are equivalent.

The alternative hypothesis: additives are not equivalent.

F-stat= 11.5333 p-value= 0.02184

Since p-value < 0.05 at the level of significance (0.05), we reject Ho

Conclusion: there is enough evidence to reject the claim that additives are equivalent

Point c: We need to increase the size of the random sample to make sure about the results and be more confident about the decisions we will take and also try to make the population be normally distributed.

Part two:

```
aggregate(problem_2.1[, 3], list(problem_2.1$Gasoline), mean)
```

```
##   Group.1 Mileage
## 1       1 127.450
```

```
## 2       2 128.325
## 3       3 128.575
```

Point a: Average value for each gasoline:

- gasoline 1: 127.450
- gasoline 2: 128.325
- gasoline 3: 128.575

```
aggregate(problem_2.1[, 3], list(problem_2.1$Additive), mean)
```

```
##   Group.1 Mileage
## 1       1 125.875
## 2       2 131.550
## 3       3 126.925
```

Point b: Average value for each Additive:

- Additive 1: 125.875
- Additive 2: 131.550
- Additive 3: 126.925

```
aggregate(problem_2.1[, 3], list(problem_2.1$Gasoline,problem_2.1$Additive),
mean)
```

```
##   Group.1 Group.2 Mileage
## 1       1       1 124.225
## 2       2       1 126.275
## 3       3       1 127.125
## 4       1       2 131.250
## 5       2       2 130.325
## 6       3       2 133.075
## 7       1       3 126.875
## 8       2       3 128.375
## 9       3       3 125.525
```

Point c: gasoline and additive combination that produced maximum value and minimum value on average:

- maximum value on average: 133.075
- minimum value on average: 124.225

```
model_3<- aov(problem_2.1$Mileage ~
problem_2.1$Gasoline*problem_2.1$Additive)
anova(model_3)
```

```
## Analysis of Variance Table
##
## Response: problem_2.1$Mileage
##                                          Df  Sum Sq Mean Sq F value
Pr(>F)
## problem_2.1$Gasoline                      2   8.375   4.187  107.68
1.359e-13
```

```
## problem_2.1$Additive                          2 218.795 109.397 2813.08 <
2.2e-16
## problem_2.1$Gasoline:problem_2.1$Additive  4  41.330  10.332  265.69 <
2.2e-16
## Residuals                                  27   1.050   0.039
##
## problem_2.1$Gasoline                       ***
## problem_2.1$Additive                       ***
## problem_2.1$Gasoline:problem_2.1$Additive ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Point d: Yes, there is an interaction effect in the data.

Results of the test:

F-stat for interactions = 265.69 p-value = 2.2e-16

From the test above, we can see that the interaction is significant.

Point e: No, the new test gives strong significant confidence to reject the null with a p-value of 1.359e-13

Point f: The large f value means something is significant, while a small p value means all your results are significant.

F-stat for additives = 2813.08 p-value = 2.2e-16

The results above show that additives together have a significant effect.

Point g: A two-way ANOVA test was conducted to solve the problem above. The dataset was transformed into three variables to simplify the test.

Results of the test are below:

- F-stat for gasoline = 107.68 p-value = 1.359e-13
- F-stat for additives = 2813.08 p-value = 2.2e-16
- F-stat for interactions = 265.69 p-value = 2.2e-16

The model shows that the gasoline used does affect the mileage.

Point h: From the result, we got in point c. The best gasoline and additive combination that will produce the maximum value is gasoline 3 with additive 2.

I will choose this combination since it is the highest maximum value average and the model results show significant results.

Problem 3

Point a: We will collect sample data of the three parties based on gender. it is not necessary to be equal samples of parties (political affiliation) or gender. First, segment the

parties of political affiliation (in our example we have three parties) then filter the data based on gender.

Point b: 300 people were sampled. which is the total of all parties and gender together.

Point c: - 156 women were sampled which is the total of all parties. - 144 men were sampled which is the total of all parties.

Point d: Party number B has the greatest number of representatives

Point e: To test whether local Men and women have an association with parties A, B, and C or not.

The null hypothesis: There is no association between local people and parties A, B and C.

The alternative hypothesis: There is an association between local people and parties A, B and C.

```
expected_partA_women <- (156*120)/300
expected_partB_women <- (156*128)/300
expected_partC_women <- (156*52)/300

cat("Expected party A women: ",(156*120)/300,'\n')

## Expected party A women:  62.4

cat("Expected party B women: ",(156*128)/300,'\n')

## Expected party B women:  66.56

cat("Expected party C women: ",(156*52)/300,'\n')

## Expected party C women:  27.04
```

Point f: Females are expected to be in parties:

- A: 62.4
- B: 66.56
- C: 27.04

```
expected_partA_men <- (144*120)/300
expected_partB_men <- (144*128)/300
expected_partC_men <- (144*52)/300

cat("Expected party A men: ",(144*120)/300,'\n')

## Expected party A men:  57.6

cat("Expected party B men: ",(144*128)/300,'\n')

## Expected party B men:  61.44

cat("Expected party C men: ",(144*52)/300,'\n')
```

```
## Expected party C men:  24.96
```

Point g: Males are expected to be in parties:

- A: 57.6
- B: 61.44
- C: 24.96

Point h: The chi-square test is a statistical test that compares observed and anticipated outcomes. The goal of this test is to figure out whether a disparity between observed and predicted data is due to chance or a link between the variables you're looking at.

Point i:

```
problem_3_TS <- ((68-62.4)^2/62.4)+ ((56-66.56)^2/66.56)+ ((32-
27.04)^2/27.04)+((52-57.6)^2/57.6)+
  ((72-61.44)^2/61.44) +((20-24.96)^2/24.96)
cat("T-test:" ,problem_3_TS)

## T-test: 6.432857
```

Since (r − 1)(s − 1) the degree of freedom = 2

```
problem_3_Pvalue <- 1-pchisq(problem_3_TS,2)
cat("p_value: " ,problem_3_Pvalue)

## p_value:  0.04009802
```

Point j: Since p-value (0.04) < 0.05, the null hypothesis is rejected at the 5 percent level of significance. That is, the hypothesis that gender and political affiliation of members of the population are independent is rejected at the 5 percent level of significance.

References:

Field, A. P., Miles, J., & Field Zoë. (2017). Discovering statistics using R. W. Ross MacDonald School Resource Services Library.

Ross, S. M. (2021). Introduction to probability and statistics for engineers and scientists. Academic Press, an imprint of Elsevier.

Peck, R., Short, T., & Olsen, C. (2020). Introduction to statistics and data analysis. Cengage.