Lab Sheet-9

Statistical Analysis

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). We describe correlations with a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by *r*

- The closer *r* is to zero, the weaker the linear relationship.
- Positive *r* values indicate a positive correlation, where the values of both variables tend to increase together.
- Negative *r* values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.

```
s1=select(mtcars, gear, carb)

s1

cor(s1)
```

```
cor(s1, use = "complete.obs")
```

Proportion tests

`prop.test()` can be used for testing the null that the proportions (probabilities of success) in several groups are the same, or that they equal certain given values.

```
prop.test(c(15,25),c(100,100))

prop.test(c(45,66),c(100,110))

Null hypothesis: The two proportions are statistically same

Alternative Hypothesis: the two proportions are statistically different

If p value<0.05, reject Null Hypothesis
```

Chi-Square test is a statistical method to determine if two categorical variables have a significant correlation between them. Both those variables should be from same population and they should be categorical like − Yes/No, Male/Female, Red/Green etc.

Syntax

The function used for performing chi-Square test is **chisq.test()**.

The basic syntax for creating a chi-square test in R is −

chisq.test(data)

Following is the description of the parameters used −

- **data** is the data in form of a table containing the count value of the variables in the observation.

Example

We will take the Cars93 data in the "MASS" library which represents the sales of different models of car in the year 1993.

For our model we will consider the variables "AirBags" and "Type". Here we aim to find out any significant correlation between the types of car sold and the type of Air bags it has. If correlation is observed we can estimate which types of cars can sell better with what types of air bags.

```
# Load the library.
library("MASS")

# Create a data frame from the main data set.
car.data <- data.frame(Cars93$AirBags, Cars93$Type)

# Create a table with the needed variables.
cardata = table(Cars93$AirBags, Cars93$Type)
print(cardata)

# Perform the Chi-Square test.
print(chisq.test(cardata))
```

When we execute the above code, it produces the following result −

```
                    Compact Large Midsize Small Sporty Van
  Driver & Passenger       2     4       7     0      3   0
  Driver only              9     7      11     5      8   3
  None                     5     0       4    16      3   6
```

```
	Pearson's Chi-squared test

data:  car.data
X-squared = 33.001, df = 10, p-value = 0.0002723

Warning message:
In chisq.test(car.data) : Chi-squared approximation may be incorrect
```

Conclusion:

Null Hypothesis: the variables are independent

Alternative Hypothesis: the variables are dependent

P<0.05: reject null hypothesis and accept A.H

The result shows the p-value of less than 0.05 which indicates a strong correlation.

Fisher’s Exact Test

Independence tests are used to determine if there is a significant relationship between two categorical variables. There exists two different types of independence test:

- the Chi-square test (the most common)
- the Fisher’s exact test

On the one hand, the Chi-square test is used when the sample is large enough. On the other hand, the Fisher’s exact test is used when the sample is small

Hypotheses

The hypotheses of the Fisher’s exact test are the same than for the Chi-square test, that is:

- H_0 : the variables are independent, there is **no** relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable
- H_1 : the variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable

Example

Data

For our example, we want to determine whether there is a statistically significant association between smoking and being a professional athlete. Smoking can only be "yes" or "no" and being a professional athlete can only be "yes" or "no". The two variables of interest are qualitative variables and we collected data on 14 persons.

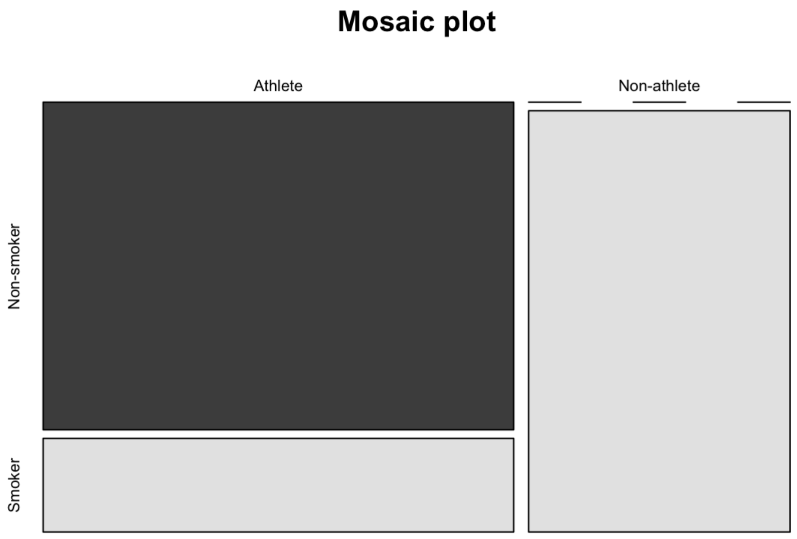
Observed frequencies

Our data are summarized in the contingency table below reporting the number of people in each subgroup:

```
dat <- data.frame(
  "smoke_no" = c(7, 0),
  "smoke_yes" = c(2, 5),
  row.names = c("Athlete", "Non-athlete"),
  stringsAsFactors = FALSE
)
colnames(dat) <- c("Non-smoker", "Smoker")


dat
##             Non-smoker Smoker
## Athlete              7      2
## Non-athlete          0      5
```

It is also a good practice to draw a mosaic plot to visually represent the data:

```
mosaicplot(dat,
  main = "Mosaic plot",
  color = TRUE
)
```



We can already see from the plot that the proportion of smokers in the sample is higher among non-athletes than athlete. The plot is however not sufficient to conclude that there is such a significant association in the population.

Expected frequencies

Remember that the Fisher's exact test is used when there is at least one cell in the contingency table of the expected frequencies below 5. To retrieve the expected frequencies, use the `chisq.test()` function together with `$expected`:

```
chisq.test(dat)$expected
## Warning in chisq.test(dat): Chi-squared approximation may be incorrect
##             Non-smoker Smoker
## Athlete            4.5    4.5
## Non-athlete        2.5    2.5
```

The contingency table above confirms that we should use the Fisher's exact test instead of the Chi-square test because there is at least one cell below 5.

Fisher's exact test in R

To perform the Fisher's exact test in R, use the `fisher.test()` function as you would do for the Chi-square test:

```
test <- fisher.test(dat)
test


##  Fisher's Exact Test for Count Data
##
## data:  dat
## p-value = 0.02098
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.449481      Inf
## sample estimates:
## odds ratio
##        Inf
```

The most important in the output is the p-value. You can also retrieve the p-value with:

```
test$p.value
## [1] 0.02097902
```

Note that if your data is not already presented as a contingency table, you can simply use the following code:

```
fisher.test(table(dat$variable1, dat$variable2))
```

where `dat` is the name of your dataset, `variable1` and `variable2` correspond to the names of the two variables of interest.

Conclusion and interpretation

From the output and from `test$p.value` we see that the p-value is less than the significance level of 5%. Like any other statistical test, if the p-value is less than the significance level, we can reject the null hypothesis.

⇒⇒ In our context, rejecting the null hypothesis for the Fisher’s exact test of independence means that there is a significant relationship between the two categorical variables (smoking habits and being an athlete or not). Therefore, knowing the value of one variable helps to predict the value of the other variable.

ANOVA Test

ANOVA also known as Analysis of variance is used to investigate relations between categorical variables and continuous variable in R Programming. It is a type of hypothesis testing for population variance.

R – ANOVA Test

ANOVA test involves setting up:

Null Hypothesis: All population means are equal.

Alternate Hypothesis: Atleast one population mean is different from other.

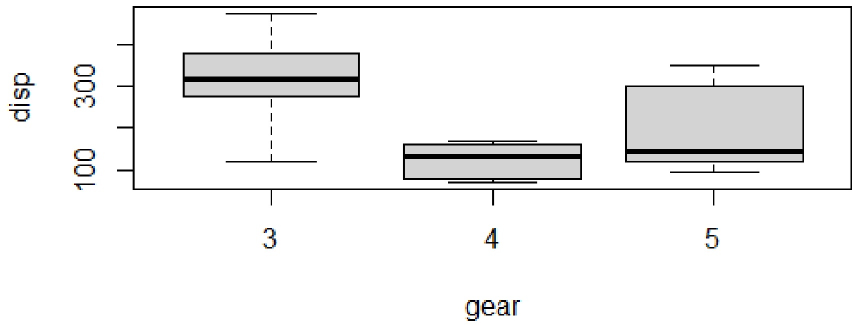
ANOVA tests are of two types:

One way ANOVA: It takes one categorical group into consideration.

Two way ANOVA: It takes two categorical group into consideration.

Performing One Way ANOVA test

```
# Installing the package
install.packages("dplyr")

# Loading the package
library(dplyr)

# Variance in mean within group and between group
boxplot(mtcars$disp~factor(mtcars$gear),xlab = "gear", ylab =
"disp")

# Step 1: Setup Null Hypothesis and Alternate Hypothesis
# H0 = mu = mu01 = mu02(There is no difference between average
displacement for different gear)
# H1 = Not all means are equal
# Step 2: Calculate test statistics using aov function
mtcars_aov <-
aov(mtcars$disp~factor(mtcars$gear))
summary(mtcars_aov)

# Step 3: Calculate F-Critical Value
# For 0.05 Significant value, critical value = alpha = 0.05

# Step 4: Compare test statistics with F-Critical value
# and conclude test p < alpha, Reject Null Hypothesis
```



The box plot shows the mean values of gear with respect of displacement. Here, categorical variable is gear on which factor function is used and continuous variable is disp.

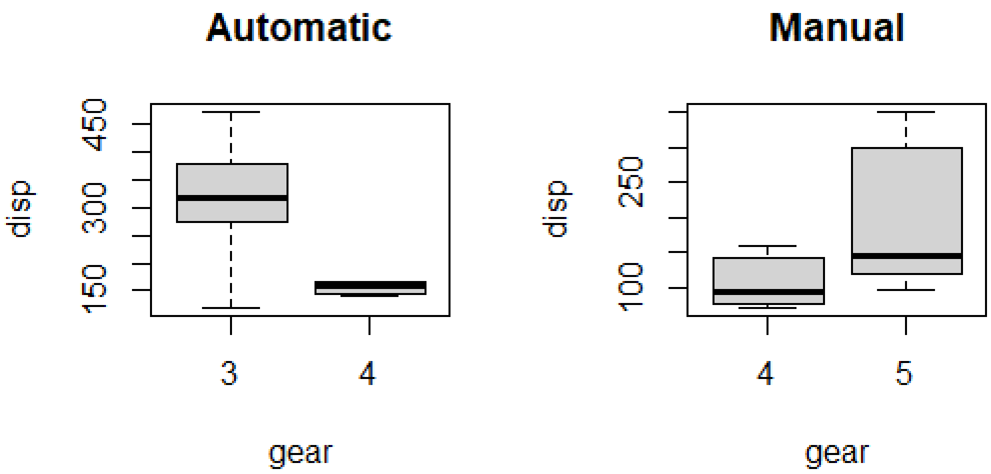
The summary shows that the gear attribute is very significant to displacement(Three stars denoting it). Also, the P value is less than 0.05, so proves that gear is significant to displacement i.e related to each other and we reject the Null Hypothesis.

Performing Two Way ANOVA test in R

Two-way ANOVA test is performed using mtcars dataset which comes preinstalled with dplyr package between disp attribute, a continuous attribute and gear attribute, a categorical attribute, am attribute, a categorical attribute.

```
# Installing the package
install.packages("dplyr")

# Loading the package
library(dplyr)
```

```
# Variance in mean within group and between group
boxplot(mtcars$disp~mtcars$gear, subset = (mtcars$am == 0),
        xlab = "gear", ylab = "disp", main = "Automatic")
boxplot(mtcars$disp~mtcars$gear, subset = (mtcars$am == 1),
            xlab = "gear", ylab = "disp", main = "Manual")

# Step 1: Setup Null Hypothesis and Alternate Hypothesis
# H0 = mu0 = mu01 = mu02(There is no difference between
# average displacement for different gear)
# H1 = Not all means are equal

# Step 2: Calculate test statistics using aov function
mtcars_aov2 <- aov(mtcars$disp~factor(mtcars$gear) *
                            factor(mtcars$am))
summary(mtcars_aov2)

# Step 3: Calculate F-Critical Value
# For 0.05 Significant value, critical value = alpha = 0.05

# Step 4: Compare test statistics with F-Critical value
# and conclude test p < alpha, Reject Null Hypothesis
```



The box plot shows the mean values of gear with respect to displacement. Here, categorical variables are gear and am on which factor function is used and continuous variable is disp.

The summary shows that the gear attribute is very significant to displacement (Three stars denoting it) and am attribute is not much significant to displacement. P-value of gear is less than 0.05, so it proves that gear is significant to displacement i.e related to each other. P-value of am is greater than 0.05, am is not significant to displacement i.e not related to each other.

Results

We see significant results from boxplots and summaries.

- Displacement is strongly related to Gears in cars i.e displacement is dependent on gears with $p < 0.05$.
- Displacement is strongly related to Gears but not related to transmission mode in cars with p 0.05 with am.