R Lab 4: One-way Analysis of Variance (ANOVA)

1. Objectives

- ANOVA: The test
- ANOVA: The assumptions

2. One-way Analysis of Variance

The Analysis of Variance (ANOVA) has its name due to the fact that it helps to compare the variation among means of several groups with the variation within groups. ANOVA tests the null hypothesis that the population means are all equal. The alternative is that they are not all equal.

Please review the assumptions for one-way ANOVA test in your lecture notes.

Exercise 1:

If a supermarket product is frequently offered at a reduced price, do customers expect the price of the product to be lower in the future? This question was examined by researchers in a study conducted on students enrolled in an introductory management course at a large midwestern university. For 10 weeks, 160 subjects read weekly ads for the same product. Students were randomly assigned to read 1, 3, 5, or 7 ads featuring price promotions during the 10-week period. They were then asked to estimate what the product's price would be the following week. We want to test if the expected prices are different between these four groups of students. Use α = 0.05. Data are stored in **pricepromo.csv**.

Firstly, import the **pricepromo.csv** data frame into R and assign it to **exprice**.

```
➢ exprice<-read.table("pricepromo.csv",header=TRUE,sep=",",quote="\"”
  ,stringsAsFactors = FALSE)
➢ exprice
➢ str(exprice)
```

It is necessary to convert the treatment variable **Promo** into factor.

```
➢ exprice$Promo <- factor(exprice$Promo, levels =c("1", "3", "5","7")
  ,labels = c("1ad","3ads","5ads","7ads"))
➢ exprice$Promo
➢ table(exprice$Promo)
```

We can find the mean of **expected price** for each treatment groups.

```
➢ by(exprice$Price, exprice$Promo, mean)
```

We can also find the standard deviation of **expected price** for each treatment group. From the output of standard deviations, we can assess the assumption of equal variances.

```
➢ by(exprice$Price, exprice$Promo, sd)
```

The rule of thumb tells us that we need not be concerned with variance assumption violation as long as the largest standard deviation is less than **twice** the smallest.

The R output for the above **by()** function is as follows:

```
exprice$Promo: 1ad
[1] 0.3621155
------------------------------------------------------------------
exprice$Promo: 3ads
[1] 0.3406328
------------------------------------------------------------------
exprice$Promo: 5ads
[1] 0.2887124
------------------------------------------------------------------
exprice$Promo: 7ads
[1] 0.3224974
```

Is it reasonable to conclude that the variances are equal?

In case the ratio of largest over smallest SDs is around 2.5, **Levene's test** can be used to check this identical variance assumption. We need to use function **leveneTest** in the package named **car**.

- `leveneTest(y, group, center = median, ...)`

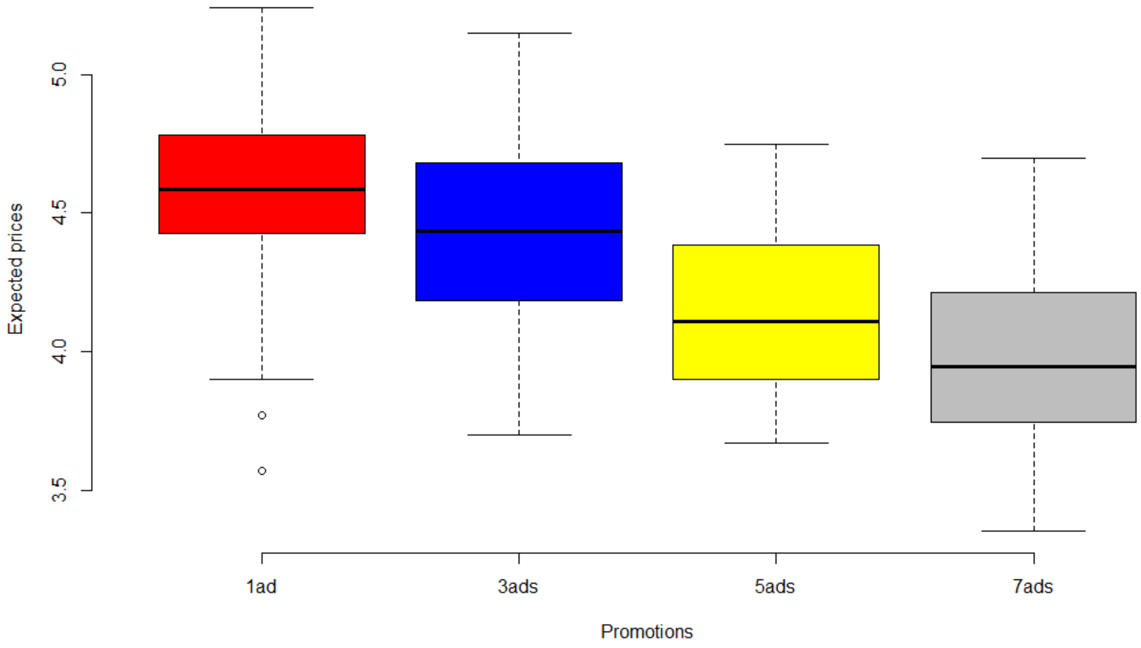
`where y: response variable; group: factor defining groups; and center = median would provide more robust test.`

Small p-value would support a conclusion that variances are different.

If the ratio is **greater than 3**, other tests should be employed in replacement of the ANOVA test.

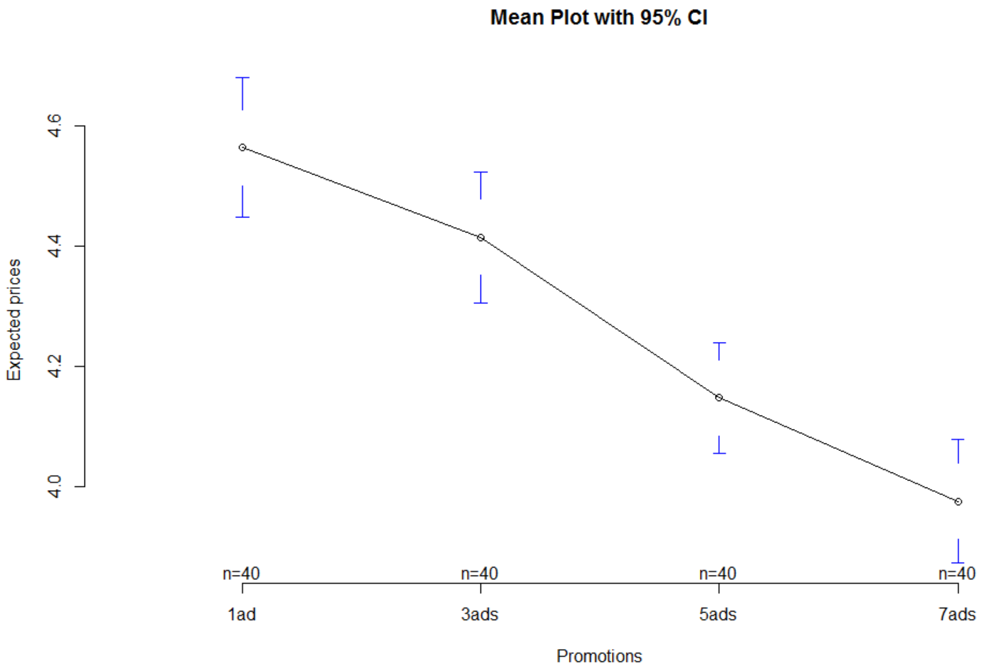
We can also use boxplots and mean plots to examine our data visually:

- `boxplot(Price ~ Promo, data = exprice, xlab = "Promotions", ylab = "Expected prices", col = c("red", "blue", "yellow","grey"))`



To draw a mean plot, we use **plotmeans** command from **gplots** package. Mean plots are drawn as the ANOVA compares means while boxplots only display medians. For distributions that are nearly symmetric, these two measures of central location will be close together.

```
➢ install.packages("gplots")
➢ library(gplots)
➢ plotmeans(Price ~ Promo, data = exprice, xlab = "Promotions", ylab
  = "Expected prices", main="Mean Plot with 95% CI")
```



We can look at the boxplots to compare within-group variation. What can small or large within-group variation tell us? Boxplots can also tell us about the **distribution of different groups** and **help to detect outliers**. When sample size is small (less than 20 for one of the groups), it is recommended that other graphical methods be used.

The ANOVA command is as follows:

```
➢ aov(y~x, data =    )
```

where **y** and **x** are the dependent and independent variables, respectively. The final argument is the name of the data structure being analyzed. The results of the ANOVA can be seen with the **summary** command.

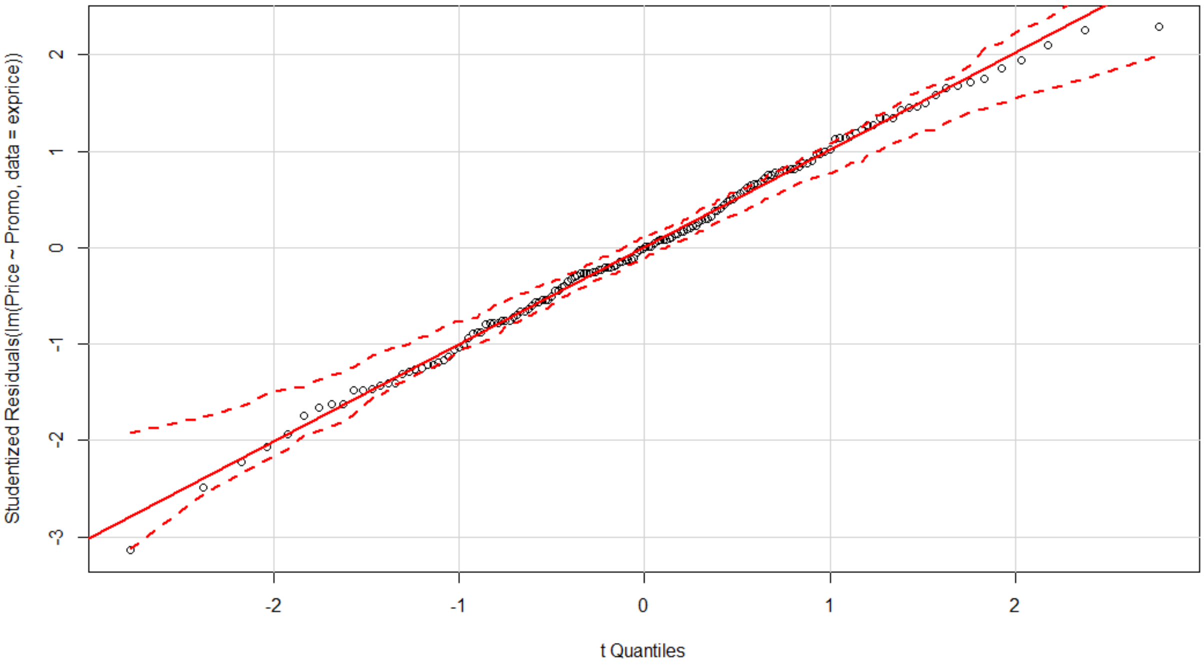
Give the hypotheses and run the one-way ANOVA with **expected price** as outcome variable and **promo** as the factor.

```
➢ aov1 <- aov(Price ~ Promo, data = exprice)
➢ summary(aov1)
             Df Sum Sq Mean Sq F value   Pr(>F)    
Promo         3  8.361  2.7868   25.66 1.52e-13 ***
Residuals   156 16.946  0.1086                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What is the test statistic with its degrees of freedom? What is the F-value? p-value? What is your conclusion about the differences between the treatment groups?

We can check the normality assumption graphically via Q-Q plot. We can draw individual Q-Q plot for each sample and check if all samples are normally distributed. Alternatively, we can draw one plot to check the normality of residuals as follows. Install the **car** package to use the **qqPlot** function.

```
➢ install.packages("car")
➢ library(car)
➢ qqPlot(lm(Price ~ Promo, data = exprice), simulate = T, labels=F)
```



Are the residuals normally distributed?

Note: If the assumptions of the ANOVA test are not met, we can use a non-parametric Kruskal-Wallis rank sum test (to be taught later in this course).

Exercise 2. Your company markets educational materials aimed at parents of young children. You are planning a new product that is designed to improve children's reading comprehension. Your product is based on new ideas from educational research, and you would like to claim that children will acquire better reading comprehension skills utilizing these new ideas than with the traditional approach. Your marketing material will include the results of a study conducted to compare two versions of the new approach with the traditional method. The standard method is called **Basal**, and the two variations of the new method are called **DRTA** and **Strat**. Education researchers randomly divided 66 children into three groups of 22. Each group was taught by one of the three methods. The response variable is a measure of reading comprehension called **COMP** that was obtained by a test taken after the instruction was completed. Can you claim that the new methods are superior to **Basal**? Use $\alpha = .05$. The data file is **eduproduct.csv**.

a) Summarize the data with mean and standard deviation for each group.
b) Is the assumption of equal standard deviations reasonable here? Explain why or why not.
c) Draw a Q-Q plot for the data in each of the three treatment groups. Summarize the information in the plots and draw a conclusion regarding the normality of these data.
d) Carry out a one-way ANOVA. Give the hypotheses, the test statistic with its degrees of freedom, and the p-value. State your conclusion.

The ouput is given below.

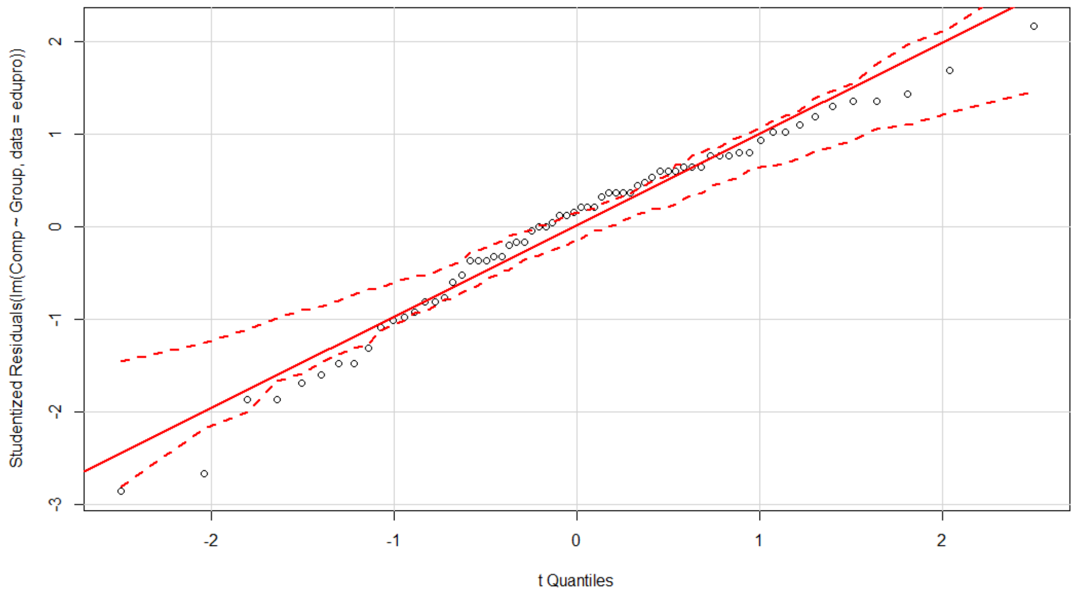
Mean for each group

```
edupro$Group: Basal
[1] 41.04545
------------------------------------------------------------------
edupro$Group: DRTA
[1] 46.72727
------------------------------------------------------------------
edupro$Group: Strat
[1] 44.27273
```

Standard deviation for each group

```
edupro$Group: Basal
[1] 5.635578
------------------------------------------------------------------
edupro$Group: DRTA
[1] 7.38842
------------------------------------------------------------------
edupro$Group: Strat
[1] 5.76675
```

Q-Q plot for the data



ANOVA output

```
            Df Sum Sq Mean Sq F value Pr(>F)
Group        2  357.3  178.65   4.481 0.0152 *
Residuals   63 2511.7   39.87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```