Two-way factorial ANOVA

1. Objectives

- Check assumptions for two-way factorial ANOVA test.
- Conduct two-way factorial ANOVA test.
- Distinguish one-way and two-way ANOVA tests.

2. Procedure

We still base on codes used for one-way ANOVA to carry out two-way ANOVA test as follows:

- `modelName <- aov(outcomevar~factor1*factor2,data=dataframe)#to see variation by interaction`
- `summary(modelName)`

Remember that this code just works in case all the **sample sizes are equal**. Besides, you must revise some techniques to check the assumptions for ANOVA such as Q-Q plot, Levene's Test for homogeneity of variance, plot to see the validity of interaction between two factors, etc.

3. Exercises

Exercise 1. Does the frequency with which a supermarket product is offered at a discount affect the price that customers expect to pay for the product? Does the percent reduction also affect this expectation? These questions were 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 received information about the products. The treatment conditions corresponded to the number of promotions (1, 3, 5, or 7) during this 10-week period and the percent at which the product was discounted (10%, 20%, 30%, and 40%). Ten students were randomly assigned to each of the 4×4=16 treatments. For our case study we will examine the data for two levels of promotions (1 and 5) and two levels of discount (10% and 30%). Data is stored in **freqdisc2.csv** file.

(a) Create a table summarising the sample size, mean, and standard deviation for each of the promotion-by-discount combinations. Is it reasonable to pool the variances? Are normality assumptions satisfied?

(b) Run the analysis of variance. Report the F statistics with degrees of freedom and p-values for each of the main effects and the interaction. What can you conclude? Write a short paragraph summarizing the results of your analysis.

Import data from **freqdisc2.csv** file into R:

- `freqdisc2 <-read.table("freqdisc2.csv", header=TRUE, sep = ",", stringsAsFactors = FALSE)`
- `str(freqdisc2)`

Because we want to see the combination promotions-by-discount, we must change variable **`Promotions`** and **`Discount`** into factors:

- `freqdisc2$Promotions <- factor(freqdisc2$Promotions, levels=c("1","5"), labels=c("1 promotion","5 promotions"))`
- `freqdisc2$Discount <- factor(freqdisc2$Discount,levels = c("10","30"), labels=c("10%","30%"))`

A crosstabulation table between **`Promotions`** and **`Discount`** variables would give you the sample size for each stratum.

- `table(freqdisc2$Promotions,freqdisc2$Discount)`

```
             10% 30%
  1promotion   10  10
  5promotions  10  10
```

To describe mean and standard deviation of **`Price`** in terms of **`Promotions`** and **`Discount`**, use the code:

- `by(freqdisc2$Price,list(freqdisc2$Promotions,freqdisc2$Discount), mean)`

You will get the following output.

```
:1promotion
: 10%
[1] 4.92
------------------------------------------------
: 5promotions
: 10%
[1] 4.393
------------------------------------------------
: 1promotion
: 30%
[1] 4.225
------------------------------------------------
: 5promotions
: 30%
[1] 3.89
```

Below is the code to get the standard deviations for each sample:

- `by(freqdisc2$Price,list(freqdisc2$Promotions,freqdisc2$Discount),sd)`

```
: 1promotion
: 10%
[1] 0.1520234
------------------------------------------------
: 5promotions
: 10%
[1] 0.2685372
------------------------------------------------
: 1promotion
: 30%
[1] 0.3856092
------------------------------------------------
: 5promotions
: 30%
[1] 0.1628906
```

Next, we're going to check the assumption of equal standard deviations. The ratio of largest SD over smallest SD is around 2.54 (which is between 2 and 3 and in this case it is not so clear to pool variances), then it's good to check again using Levene's test:

- `leveneTest(freqdisc2$Price,interaction(freqdisc2$Promotions, freqdisc2$Discount),center=median)`

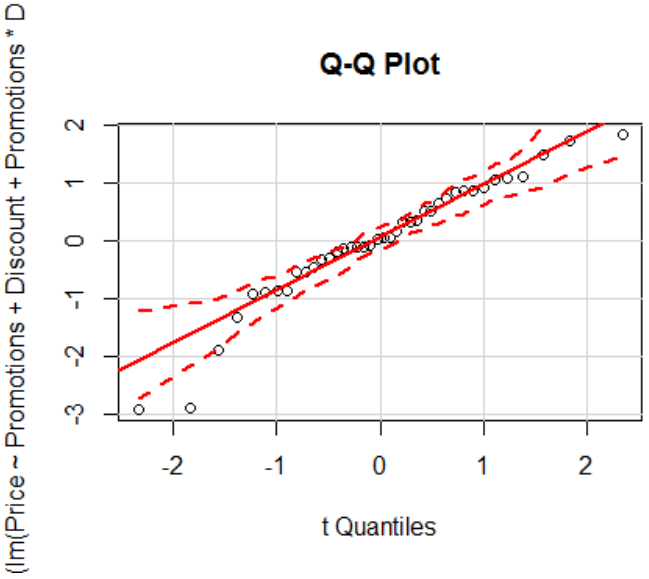
The test gives you:

```
Levene's Test for Homogeneity of Var
iance (center = median)
      Df F value  Pr(>F)
group  3  2.7878 0.05451 .
      36
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0
.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What is your conclusion about the assumption of equal standard deviations?

We check the assumption of normality using Q-Q plot:

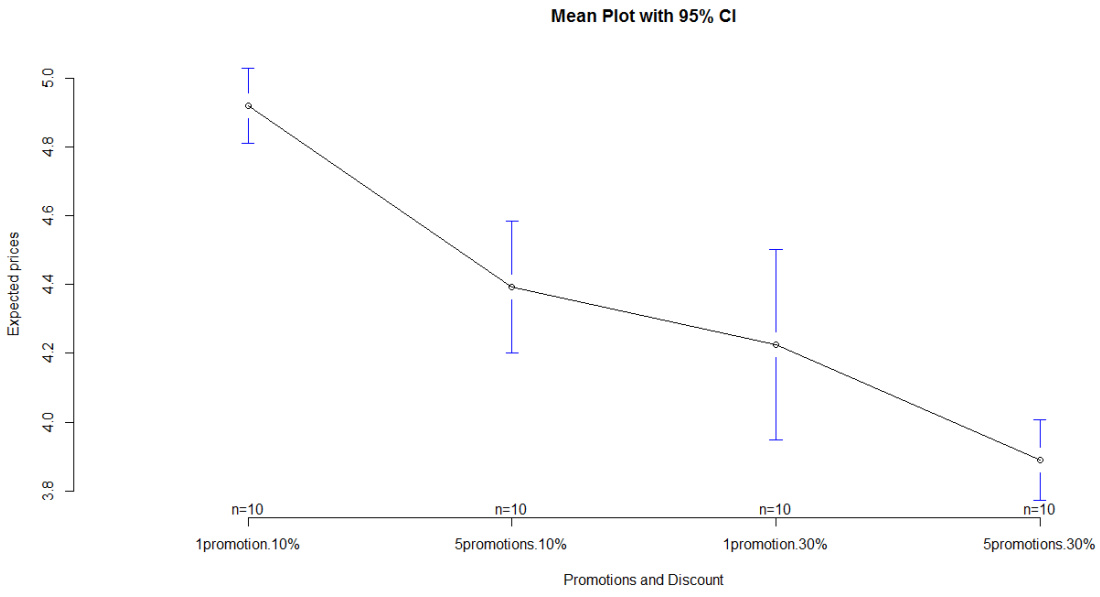
- `library(car)`
- `qqPlot(lm(Price ~ Promotions + Discount + Promotions*Discount, data=freqdisc2), simulate=T, main="Q-Q Plot", labels=F)`



What can you say about normality of residuals based on the above Q-Q plot?

The sample sizes of all groups are not so large (just 10 observations) then it's not appropriate to use boxplots to compare 4 groups, instead, we'd like to use meanplots. The codes and outputs are provided below:

- `install.packages("gplots")`
- `library(gplots)`
- `plotmeans(Price ~ interaction(Promotions,Discount), data = freqdisc2, xlab = "Promotions and Discount", ylab = "Expected prices", main="Mean Plot with 95% CI")`



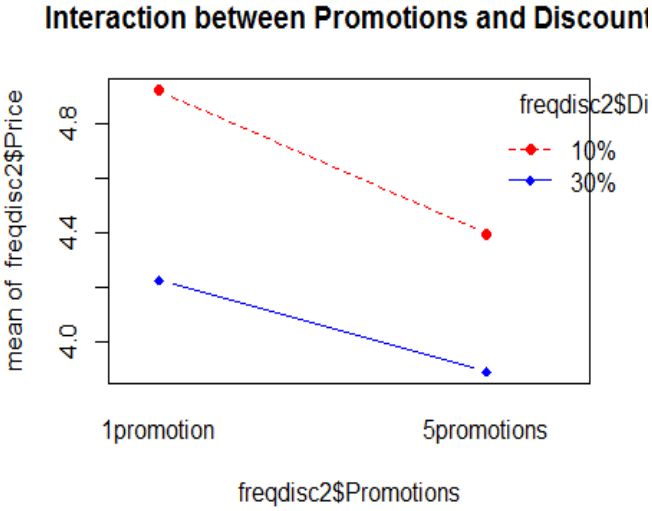
Two-way ANOVA

Now we're going to run two-way ANOVA test with **Price** as outcome variable and **Promotions** and **Discounts** as two factors. We're also interested in the main effects of **Promotions** and **Discount** and their ***interaction***, so we use the format `Price ~ Promotions*Discount`:

- `freqdisc2.result<-aov(Price ~ Promotions*Discount, data = freqdisc2)`
- `summary(freqdisc2.result)`

Here is the R output for two-way ANOVA test:

```
                   Df Sum Sq Mean Sq F value   Pr(>F)
Promotions          1  1.858   1.858  27.474 7.17e-06 ***
Discount            1  3.588   3.588  53.067 1.39e-08 ***
Promotions:Discount 1  0.092   0.092   1.363    0.251
Residuals          36  2.434   0.068
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Questions: How do promotions and discount and their interaction affect the expected price? Provide comments.

Interaction Plot:

We want to see the interaction between two factors graphically, so we use the `interaction.plot` function as follows.

- `interaction.plot(freqdisc2$Promotions, freqdisc2$Discount, freqdisc2$Price, type="b", col=c("red", "blue"), pch=c(16, 18), main = "Interaction between Promotions and Discount")`

Note that because the interaction effect is not significant, we do not interpret the interaction plot. In practice, you do not need to produce an interaction plot if the interaction effect is not significant.

Exercise 2. A study of cardiovascular risk factors compared runners who averaged at least 15 miles per week with a control group described as "generally sedentary." Both men and women were included in the study. The data set was constructed based on information provided in P. D. Wood et al., "Plasma lipoprotein distributions in male and female runners," in P. Milvey (ed.), The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies, New York Academy of Sciences, 1977. The study design is a 2×2 ANOVA with the factors **group and gender**. There were 200 subjects in each of the four combinations. The variables are ID, a numeric subject identifier; Group, with values "Control" and "Runners"; Gender, with values Female and Male; and HeartRate, heart rate after the subject ran for six minutes on a treadmill. Analyze the data using a two-way ANOVA. Summarize your findings in a short report. The data file is **runners.csv**.

1. **Import data from runners.csv into R, then check the first 6 subjects (using `head()`) as well as structure of this dataframe.**

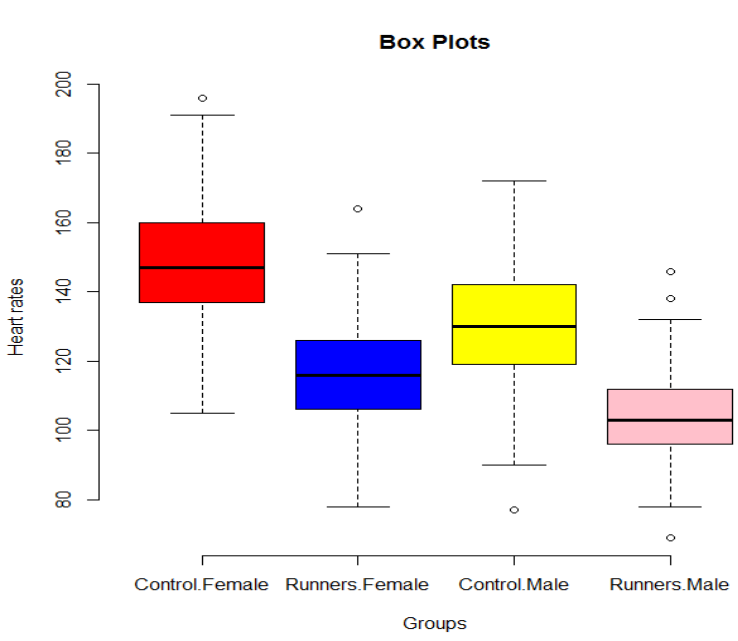
```
  Id   Group Gender HeartRate
1  1 Control Female       159
2  2 Control Female       183
3  3 Control Female       140
4  4 Control Female       140
5  5 Control Female       125
6  6 Control Female       155

'data.frame': 800 obs. of  4 variables:
 $ Id       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Group    : Factor w/ 2 levels "Control","Runners": 1 1 1 1 1 1 1 1 1 1 ...
 $ Gender   : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ HeartRate: int  159 183 140 140 125 155 148 132 158 136 ...
```

2. **Crosstabulation table between 2 factors:**

```
         Female Male
 Control    200  200
 Runners    200  200
```

Graphical description:



3. **Means for groups:**

```
: Control
: Female
[1] 148
------------------------------------------------
: Runners
: Female
[1] 115.985
------------------------------------------------
: Control
: Male
[1] 130
------------------------------------------------
```

```
: Runners
: Male
[1] 103.975
```

4. **Standard deviation for groups:**

```
: Control
: Female
[1] 16.27095
------------------------------------------------
: Runners
: Female
[1] 15.97154
------------------------------------------------
: Control
: Male
[1] 17.10035
------------------------------------------------
: Runners
: Male
[1] 12.49942
```

5. **Check the homogeneity of variances:**

Check the assumption of equal standard deviations using the rule we learnt in the lecture. Be careful when you use the Levene's test for large sample sizes:

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   3  5.7339 0.0006971 ***
      796
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. **Check the normality of residuals:**

7. **Two-way ANOVA:**

```
              Df Sum Sq Mean Sq F value  Pr(>F)
Group          1 168432  168432 695.647 < 2e-16 ***
Gender         1  45030   45030 185.980 < 2e-16 ***
Group:Gender   1   1794    1794   7.409 0.00663 **
Residuals    796 192730     242
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note: If the interaction effect is significant, you should ignore the main effects.

Interaction Plot