**Statistical Analysis**

Master in Statistics and Information Management

Professor *Ana Cristina Costa* (ccosta@novaims.unl.pt)

# HOMEWORK

## LU7: Analysis of Variance (ANOVA)

**In the following exercises, and if necessary, consider that the necessary assumptions for the application of ANOVA are verified. Consider the 5% level of significance, except otherwise stated.**

**1.** Consider the following ANOVA table with data that allow to test the equality of k population means.

| Source of variation | Degrees of freedom | Sum of squares | Mean squares | F |
|---|---|---|---|---|
| ***Treatments*** | 3 | 6.159 | | |
| ***Error*** | 20 | | | |
| ***Total*** | | 16.619 | | |

**a)** Complete the ANOVA table.

**b)** What assumptions are required in using the one-way ANOVA?

**c)** Formulate the hypotheses to be tested.

**d)** For a 5% significance level, which should be the decision rule? And the decision?

**2.** In order to verify if the average age of women attending the 1st family planning services was the same in four different cities (A, B, C and D), a sample of the services' clients was collected for each one of the cities. The following results were obtained: SST=130.964 ; MSTr=24.321 and also

| | City A | City B | City C | City D |
|---|---|---|---|---|
| *Sample size* | 6 | 8 | 7 | 7 |
| *Sample mean* | 19 | 18 | 22 | 21 |
| *Sample variance* | 1.3 | 1.7 | 4.3 | 2.3 |

**a)** Determine the ANOVA table.

**b)** Test the ANOVA hypotheses.

**3.** The data of the following table refers to the quotation of four different groups of shares.

| | **Quotations of different groups of shares** | | | |
|---|---|---|---|---|
| **Observations** | **1** | **2** | **3** | **4** |
| 1 | 15.1 | 14.9 | 15.4 | 15.6 |
| 2 | 15.0 | 15.2 | 15.2 | 15.5 |
| 3 | 14.9 | 14.9 | 16.1 | 15.8 |
| 4 | 15.7 | 14.8 | 15.3 | 15.3 |
| 5 | 15.4 | 14.9 | 15.2 | 15.7 |
| 6 | 15.1 | 15.3 | 15.2 | 15.7 |
| **Sample means** | **15.2** | **15.0** | **15.4** | **15.6** |
| **Sample variances** | **0.088** | **0.040** | **0.124** | **0.032** |

**a)** Test the hypothesis of homogeneity of the variances among the four groups of shares using the Bartlett's test.

**b)** Represent the results in the ANOVA table and check if there are real differences between the four groups of shares.

**c)** Determine the p-value of the F-test.

**d)** If there are significant differences between the groups, verify which groups differ significantly among themselves.

**4.** With the objective of comparing the housing prices of four different cities, five houses that were for sale in each of the cities were randomly selected and their prices registered (in thousands of euros). Is there any evidence that the four cities differ significantly in relation to the average housing price? If yes, indicate which of the cities have significantly different prices.

| | **City A** | **City B** | **City C** | **City D** |
|---|---|---|---|---|
| **Housing prices** (thousands of euros) | 110 | 72 | 88 | 57 |
| | 160 | 38 | 66 | 81 |
| | 93 | 45 | 112 | 181 |
| | 206 | 108 | 47 | 165 |
| | 171 | 42 | 52 | 106 |
| **Sample means** | **148** | **61** | **73** | **118** |
| **Sample variances** | **2126.5** | **870** | **727.5** | **2852.5** |

**5.** In a market study, whose main objective was to detect the differences in behaviour of readers of the 3 weekly newspapers (*Expresso*, *Independente* and *Semanário*), random independent samples were collected concerning the reading time (in minutes) of each reader:

*Expresso*: sample size $n_1 = 8$, sample mean = 93

*Independente*: sample size $n_2 = 6$, sample mean = 75

*Semanário*: sample size $n_3 = 6$, sample mean = 70

Admit that the variance of the reading time is equal in the three groups of readers. The table resulting from the application of the analysis of variance is the following:

| Source of variation | Degrees of freedom | Sum of squares | Mean squares | F |
|---|---|---|---|---|
| ***Treatments*** | | 2092.2 | | |
| ***Error*** | | 3118 | | |
| ***Total*** | | 5210.2 | | |

**a)** Complete the ANOVA table.

**b)** Verify that the mean reading times are identical in the three populations (readers of the *Expresso*, the *Independente* and the *Semanário*).

**c)** Which differ significantly?

## Solutions

Question 1)

a) $F_{obs}$ = 3.9254

d) Reject $H_0$

Question 2)

a) $F_{obs}$ = 10,064

b) Reject $H_0$

Question 3)

a) $Q_{obs}$ = 2.7622; $Q_{crit}$ = 7.8147 (p-value = 0.4298)

b) $F_{obs}$=5.6338; Reject $H_0$

c) 0.0058

d) Reject $H_0$: $\mu_2 = \mu_4$

Question 4) Barlett's test: $Q_{obs}$ = 2.3314; $Q_{crit}$ = 7.8147 (p-value= 0.5065)
ANOVA: $F_{obs}$ = 4.9449; p-value= 0.0129
Tuckey's HSD test: reject $H_0$: $\mu_A = \mu_B$ and $H_0$: $\mu_A = \mu_C$

Question 5)

a) $F_{obs}$ = 5.7036

b) Reject $H_0$

c) Reject $H_0$: $\mu_1 = \mu_3$

# APPENDIX

## DEMONSTRATIONS

Assuming that the One-way ANOVA assumptions are verified, the following properties are demonstrated.

**Property 1: $SQE/\sigma^2 \sim \chi^2_{(n-k)}$.**

$$\frac{\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_i\right)^2}{\sigma^2}=\frac{(n_i-1)S_i^2}{\sigma^2}\sim\chi^2_{(n_i-1)}$$

$$\frac{SQE}{\sigma^2}=\sum_{i=1}^{k}\left(\frac{\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_i\right)^2}{\sigma^2}\right)\sim\chi^2_{\left(\sum_{i=1}^{k}(n_i-1)\right)}$$

$$\Rightarrow\frac{SQE}{\sigma^2}\sim\chi^2_{(n-k)}$$

**Property 2: SQE/(n–k) is an unbiased estimator of $\sigma^2$, regardless of the fact that the null hypothesis $H_0$: $\mu_1=\mu_2=\ldots=\mu_k$ is true or false.**

By *Property 1*: $\frac{SQE}{\sigma^2}\sim\chi^2_{(n-k)}\Rightarrow E\left(\frac{SQE}{\sigma^2}\right)=n-k\Leftrightarrow E\left(\frac{SQE}{n-k}\right)=\sigma^2$

Consequently, SQE/(n–k) is an unbiased estimator of $\sigma^2$, regardless of $H_0$.

**Property 3: Under the null hypothesis $H_0$: $\mu_1=\mu_2=\ldots=\mu_k$, we have $SQT/\sigma^2\sim\chi^2_{(n-1)}$.**

If $H_0$ is true then $\mu_1=\mu_2=\ldots=\mu_k$ and, therefore, the sample formed by the global set of *n* observations is a sample from a population $N(\mu,\sigma^2)$ which implies that

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}\right)^2}{\sigma^2}=\frac{SQT}{\sigma^2}\sim\chi^2_{(n-1)}$$

**Property 4: Under the null hypothesis $H_0$: $\mu_1=\mu_2=\ldots=\mu_k$, we have $SQTr/\sigma^2\sim\chi^2_{(k-1)}$.**

By previous properties we know that $SQE/\sigma^2\sim\chi^2_{(n-k)}$ and that if $H_0$ is true then $SQT/\sigma^2\sim\chi^2_{(n-1)}$. Therefore,

$$\frac{SQT}{\sigma^2}=\frac{SQTr}{\sigma^2}+\frac{SQE}{\sigma^2}\Rightarrow\frac{SQTr}{\sigma^2}\sim\chi^2_{(k-1)}$$

**<u>Property 5</u>: Under the null hypothesis $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$, we have that SQTr/(k–1) is an unbiased estimator of $\sigma^2$.**

By *Property 4*: $\frac{\text{SQTr}}{\sigma^2} \sim \chi^2_{(k-1)} \Rightarrow \text{E}\left(\frac{\text{SQTr}}{\sigma^2}\right) = k-1 \Leftrightarrow \text{E}\left(\frac{\text{SQTr}}{k-1}\right) = \sigma^2$

Consequently, SQTr/(k–1) is an unbiased estimator of $\sigma^2$ when $H_0$ is true.

**<u>Property 6</u>**: If $X \sim \chi^2_{(n)}$ and $Y \sim \chi^2_{(m)}$ are independent random variables then

$$F = \frac{X/n}{Y/m} = \frac{\chi^2_{(n)}/n}{\chi^2_{(m)}/m} \sim F_{(n,m)}$$

**<u>Property 7</u>: Under the null hypothesis $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$, the test statistic of the ANOVA F-test is**

$$F = \frac{MQTr}{MQE} \sim F_{(k-1,n-k)}$$

Considering the results of the *Properties 1, 4 and 6*, we conclude that:

$$\frac{\text{MQTr}}{\text{MQE}} = \frac{SQTr/(k-1)}{SQE/(n-k)} = \frac{SQTr/\sigma^2(k-1)}{SQE/\sigma^2(n-k)} = \frac{\chi^2_{(k-1)}/(k-1)}{\chi^2_{(n-k)}/(n-k)} \sim F_{(k-1,n-k)}$$