Ordinary session
Final exam

Time: 120 min.

- **You are not allowed to use any documentation apart from the formula sheet you have received.**
- **Use 4 decimal digits in all calculations and results.**

1. **(2 Points)** In a communication system, messages are encoded in base 2, that is, with the symbols 0 and 1. The probability of emitting a 0 is 0.7, and of emitting a 1 is 0.3. Whatever symbol is emitted, the probability of receiving that same symbol is 0.9.

   (a) **(1 Point)** Find the probability of receiving a 1.

   Solution

   E_i: the symbol i is emitted R_i: the symbol i is received

   $P(E_0) = 0.7$ $P(E_1) = 0.3$

   $P(R_0|E_0) = P(R_1|E_1) = 0.9$

   By the Total Probability Theorem:
   $P(R_1) = P(R_1|E_0)P(E_0) + P(R_1|E_1)P(E_1)$

   $P(R_1|E_0) = 1 - P(R_0|E_0) = 1 - 0.9 = 0.1$

   Then
   $P(R_1) = 0.1 \cdot 0.7 + 0.9 \cdot ., 3 = 0.34$

   (b) **(1 Point)** Find the probability that the emitted symbol was 0 if 0 was received.

   Solution

   By Bayes' Theorem:
   $P(E_0|R_0) = \dfrac{P(R_0|E_0)P(E_0)}{P(R_0)}$

   $P(R_0) = P(R_0|E_0)P(E_0) + P(R_0|E_1)P(E_1)$

   $P(R_0|E_1) = 1 - P(R_1|E_1) = 1 - 0.9 = 0.1$

   $P(R_0) = 0.9 \cdot 0.7 + 0.1 \cdot 0.3 = 0.66$

   $P(E_0|R_0) = \dfrac{0.9 \cdot 0.7}{0.66} \approx 0.9545.$

2. **(2 Points)** Suppose that the time in hours that a student spends each week practicing sports is distributed according to a random variable with a density function given by the following function:

$$f(x) = \begin{cases} k\ e^{-0.25x} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases}.$$

   (a) **(0.5 Points)** What should the value of k be for $f(x)$ to be a density function? Justify your answer.

   Solution

   Since the given function is an exponential function with the parameter 0.25, then $k = 0.25$. It can also be obtained by applying the property of the density function $\int_{-\infty}^{+\infty} f(x)dx = 1$.

   (b) **(0.5 Points)** What is the probability that the student, during a given week, dedicates more than 5 hours to sports?

   Solution

   If we denote by X the weekly time in hours dedicated by the student to sport, what they ask us is

$$P(X > 5) = 1 - P(X \leq 5) = 1 - F_X(5) = 1 - (1 - e^{-0.25 \times 5}) \approx 0.2865.$$

(c) **(1 Point)** If during 8 weeks, chosen at random and independently, the student records the weekly time dedicated to sports, what is the probability that exactly 2 of those 8 records indicate that the student has dedicated more than 5 hours to sports per week?

Solution

If we denote by Y the random variable that counts the number of weeks during which the student dedicates more than 5 hours per week to sports, then we can say that Y follows a binomial distribution with parameters $n = 8$ y $p = 0.2865$. The probability they ask us is:

$$P(Y = 2) = \binom{8}{2} 0.2865^2 (1 - 0.2865)^6 \approx 0.3032.$$

3. **(3 Points)** Two independent samples of 39 men and 35 women yielded the following means and quasi-standard deviations of the number of days that a patient remains admitted to a hospital:

$$\bar{x}_m = 7.90, \quad s_m = 6.41, \quad \bar{x}_w = 7.11, \quad s_w = 5.16.$$

(a) **(1 Point)** Calculate and interpret a 95% confidence interval for the expected number of days a woman remains in the hospital. Datum: $z_{0.025} = 1.96$.

Solution

Confidence interval for the mean of a population (large samples):

$$IC = \bar{x}_m \pm z_{\alpha/2} \sqrt{\frac{s_m^2}{35}} = [5.4005;\ 8.8195]$$

that is, the mean number of days in the hospital for women is between 5.4 and 8.8 days with a 95% confidence.

(b) **(0.5 Points)** Is it necessary to assume that the samples are large enough to answer the previous section? Justify your answer.

Solution

Clearly, the data, number of days of hospitalization, do not follow a normal distribution, since the variables are discrete and with a small range. The confidence interval is based on the Central Limit Theorem, for which we must assume that the samples are large.

(c) **(1.5 Points)** By clearly specifying the null and alternative hypotheses, can we affirm that the average number of days that a woman remains hospitalized is significantly different than the average number of days that a man remains hospitalized? Calculate and interpret the p-value of the test.

Solution

Test statement:

$$H_0 : \mu_h = \mu_m, \quad H_1 : \mu_h \neq \mu_m$$

Statistic and p-value:

$$t = \frac{7.90 - 7.11}{\sqrt{\frac{6.41^2}{39} + \frac{5.16^2}{35}}} \approx 0.5865, \qquad \text{p-valor} = 2\text{P}(Z > 0.5865) \approx 0.5552.$$

We do not reject H_0 since the p-value is very large. We conclude that there is no significant difference between the average number of days that a man remains hospitalized and the average number of days that a woman remains hospitalized.

4. **(3 Points)** We want to predict, using a multiple linear regression model, a student's grade (`Nota`) from their weekly hours of study (`Horas_Estudio`), their hours of class attendance (`Horas_Clase`) and the investment in study material (`Inversion`). The results obtained for two estimated models are shown below:

```
> modelo1 <- lm(Nota ~ Horas_Estudio + Horas_Clase + Inversion)
> summary(modelo1)

Call:
lm(formula = Nota ~ Horas_Estudio + Horas_Clase + Inversion)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47053 -0.24793 -0.00349  0.30364  1.18773

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0365743  0.2559537   4.050 0.000104 ***
Horas_Estudio  0.5009395  0.0158676  31.570  < 2e-16 ***
Horas_Clase    0.4756628  0.0610090   7.797 7.64e-12 ***
Inversion     -0.0004612  0.0043932  -0.105 0.916617
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4827 on 96 degrees of freedom
Multiple R-squared:  0.9151,Adjusted R-squared:  0.9124
F-statistic: 344.9 on 3 and 96 DF,  p-value: < 2.2e-16

> aov(modelo1)
Call:
   aov(formula = modelo1)

Terms:
                Horas_Estudio Horas_Clase Inversion Residuals
Sum of Squares      226.90966    14.16923   0.00257  22.36913
Deg. of Freedom             1           1         1        96

Residual standard error: 0.4827129
Estimated effects may be unbalanced
>
> modelo2 <- lm(Nota ~ Horas_Estudio + Horas_Clase)
> summary(modelo2)

Call:
lm(formula = Nota ~ Horas_Estudio + Horas_Clase)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47364 -0.24840 -0.00223  0.30421  1.19156

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.03447    0.25386   4.075 9.42e-05 ***
Horas_Estudio  0.50067    0.01557  32.151  < 2e-16 ***
Horas_Clase    0.47573    0.06069   7.838 5.92e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4802 on 97 degrees of freedom
Multiple R-squared:  0.9151,Adjusted R-squared:  0.9133
F-statistic: 522.6 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
> aov(modelo2)
Call:
   aov(formula = modelo2)

Terms:
                Horas_Estudio Horas_Clase Residuals
Sum of Squares      226.90966    14.16923  22.37169
Deg. of Freedom             1           1        97

Residual standard error: 0.4802458
Estimated effects may be unbalanced
```

(a) **(1 Point)** What is the best model? Justify your answer.

Solution

The second model is more suitable because all the variables are significant and the adjusted R^2 is greater.

(b) **(0.5 Points)** Write the equation of the model and interpret its coefficient of determination, R^2.

Solution

$Nota = 1.03447 + 0.50067 * Horas\_Estudio + 0.47573 Horas\_Clase.$

The model explains 91.51% of the variability observed in the notes.

(c) **(0.5 Points)** Can it be said that the errors of the models follow a normal distribution? Justify your answer.

```
> library(nortest)
> pearson.test(modelo1$residuals)

 Pearson chi-square normality test

data:  modelo1$residuals
P = 9.2, p-value = 0.5132

> pearson.test(modelo2$residuals)

 Pearson chi-square normality test

data:  modelo2$residuals
P = 8.68, p-value = 0.5627
```

Solution

In both models we cannot reject the normality hypothesis since the p-value is greater than any usual value of α.

(d) **(1 Point)** A student has invested 60€ in study material, has studied 5 hours and has attended 4 hours of class. How likely is it that she will get a grade greater than 5? Justify your answer.

Solution

As the variable `Inversion` was clearly not significant, we will use model 2 to perform the calculations: We have to `Horas_Estudio` = 5 and `Horas_Clase` = 4, then:

$$E[Nota|Horas\_Estudio = 5, Horas\_Clase = 4] = 1.03447 + 0.50067 * 5 + 0.47573 * 4 \approx 5.4407.$$

On the other hand, we have that $\widehat{\sigma} = 0.4802$, therefore, we can assume that $Nota|Horas\_Estudio = 5, Horas\_Clase = 4 \; \mathcal{N}(\mu = 5.4407, \sigma = 0.4802)$:

$$\begin{aligned} P(Nota > 5|Horas\_Estudio = 5, Horas\_Clase = 4) &= P\left(\frac{Nota - \mu}{\sigma} > \frac{5 - 5.4407}{0.4802}\right) \\ &\approx P(Z > -0.92) = 0.8212. \end{aligned}$$