ISEG – Lisbon School of Economics and Management
STATISTICS 2
Second Semester 2020/2021
Normal Exam
Monday 31 May 2021

Duration: **2 hours**

Name:
Student Number:

Question:	1	2	3	4	5	6	7	Total
Points:	20	10	30	10	50	40	40	200

Justify all your answers. A correct answer in a multiple choice question is worth 10 points; an incorrect one is worth −2.5 points. **You are allowed to use** the formula sheet, and the statistical tables that are provided on the webpage of the course, without ANY annotation. You can also use a calculator and A4 empty scrap paper. Consider a **5% significance level**, unless otherwise is stated.

1. The daily number of COVID-19 infections on the north portuguese municipalities are assumed to follow an exponential distribution with mean 35, whereas the daily number of COVID-19 infections on the south portuguese municipalities are assumed to follow an exponential distribution with mean 65. For now, assume that the COVID-19 infections, per municipality (both north and south), are independent.

(10) (a) In a sample of 10 north portuguese municipalities, what is the probability that the daily maximum number of COVID-19 infections exceeds 100?

$X_N \sim exp(\lambda = \frac{1}{35})$ $E[X] = \frac{1}{\lambda} = 35.$ $F_{X_N}(x) = 1 - e^{-\lambda x}$

$n = 10$

$$P(X_{(n)} > 100) = 1 - P(X_{(n)} < 100)$$

$$= 1 - P(X_1 < 100, \ldots, X_n < 100)$$

$$= 1 - P(X_1 < 100) * \ldots * P(X_n < 100)$$

$$= 1 - \{P(X_N < 100)\}^n$$

$$= 1 - \{F_{X_N}(100)\}^n$$

$$= 1 - (1 - e^{-\frac{1}{35} * 100})^{10}$$

$$\approx 0.4465$$

(10) (b) Consider a sample of 100 north portuguese municipalities and 100 south portuguese municipalities. Approximate the probability that the average number of daily infections in the north municipalities is at least 31 lower than the average number of daily infections in the south municipalities?

$$\bar{X}_N \underset{\text{CLT}}{\overset{\text{approx}}{\sim}} N\left(\mu=35, \frac{35^2}{100}\right) \qquad \bar{X}_S \underset{\text{CLT}}{\overset{\text{approx}}{\sim}} N\left(\mu_S=65, \frac{65^2}{100}\right)$$

$$P(\bar{X}_N < \bar{X}_S - 31) = P(\underbrace{\bar{X}_N - \bar{X}_S}_{N\left(-30, \frac{35^2+65^2}{100}\right)} < -31)$$

$$= P\left(\underbrace{\frac{\bar{X}_N - \bar{X}_S + 30}{\sqrt{(35^2+65^2)/100}}}_{N(0,1)} < \underbrace{\frac{-31+30}{\boxed{\sqrt{(35^2+65^2)/100}}}}_{-0.14}\right) \rightarrow 7.3824$$

$$= 1 - \underbrace{\Phi(0.14)}_{0.557 \text{ (tables)}} \approx 0.444$$

(10) **2**. Which of the following statements is correct (only one statement is correct)?

○ A statistic is a function of the unknown population parameter(s).

○ A statistic can assume infinitely many values, assuming that we can draw infinitely many samples, with the same sample size, from the population.

○ The repeated sampling distribution of any statistic can be approximated by the Standard Normal distribution.

☒ None of the above.

3. Let X be a random variable with probability distribution function given by:

$$f(x \mid \theta) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots, \quad 0 < \theta < 1, \quad \mathrm{E}(X) = \frac{1}{\theta} \text{ and } \mathrm{Var}(X) = \frac{1-\theta}{\theta^2}$$

(15) (a) Obtain the maximum likelihood estimator for θ.

1) $f_X(x) = \theta(1-\theta)^{x-1}$

2) $\ell_i(\theta) = \log(\theta) + (x_i - 1)\log(1-\theta)$

$\ell_n(\theta) = n\log(\theta) + \left(\sum_{i=1}^{n} x_i - n\right)\log(1-\theta)$

3) $\dfrac{d\ell_n(\theta)}{d\theta} = \dfrac{n}{\theta} - \dfrac{(\sum x_i - n)}{1-\theta} = 0$

$\Leftrightarrow \dfrac{\boxed{n(1-\theta) - \theta\sum x_i + n\theta}}{\boxed{\theta(1-\theta)}} = 0$ $\rightarrow n - \cancel{n}\theta - \theta\sum x_i + \cancel{n}\theta$ $\quad \boxed{\theta(1-\theta)} > 0$

$\Leftrightarrow n = \theta\sum_{i=1}^{n} X_i \Leftrightarrow \hat{\theta}_{ml} = \dfrac{1}{\bar{X}}$

(15) (b) Obtain the maximum likelihood estimator for $\mathrm{E}(X)$. Make sure to explain your answer.

$E[X] = \dfrac{1}{\theta}$

$\widehat{E[X]}_{ml} = \dfrac{1}{\hat{\theta}_{ml}}$ by invariance of ml to (nonlinear) transformations

$= \dfrac{1}{(1/\bar{X})} = \bar{X}.$

(10) **4.** For any random variable X, with a given probability distribution function, $f(x \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ can be a vector of unknown parameters, which of the following statements is correct (only one statement is correct)?

○ The Method of Moments estimator is always equal to the Maximum Likelihood estimator, for some unknown parameter of $\boldsymbol{\theta}$.

○ There cannot be any estimator that takes a value below the Cramer-Rao lower bound.

○ If $X \sim \mathrm{N}(\mu, \sigma^2)$, then the Maximum Likelihood estimator for both μ and σ^2 is unbiased.

☒ None of the above.

5. Current evidence on COVID-19 shows that the time (in days) until an individual develops disease symptoms, after getting infected, follows a Normal distribution. More precisely, for male individuals, it follows a Normal distribution with unknown mean μ_M and unknown variance σ_M^2, whereas for female individuals, it follows a Normal distribution with unknown mean μ_F and unknown variance σ_F^2. The following sample results were obtained:

Male individuals: $n_M = 13 \rightarrow 11$, $\bar{x}_M = 6.257$, $s_M^2 = 0.709 \rightarrow S'^2_m$

Female individuals: $n_F = 12 \rightarrow 10$, $\bar{x}_F = 2.945$, $s_F^2 = 0.518 \rightarrow S'^2_F$

(20) (a) Based on a 99% confidence interval, comment on the statement: "in fact, the mean values for the distributions of the male and female individuals can be assumed to be the same".

step 1) Find pivotal quantity:

$$\frac{\overset{3.312}{(\bar{x}_m - \bar{x}_F)} - (\mu_m - \mu_F)}{\underset{\cancel{0.929}\ 0.341}{\sqrt{S'^2_m/11 + S'^2_F/10}}} \overset{\text{approx}}{\sim} N(0,1)$$

$\cancel{t_{n_m+n_F-2}}$ 19

$N(0,1)$

step 2) Find quantiles: $q_{0.995} = 2.576$ (tabelas)

step 3) Rewrite $P(-2.576 < PQ < 2.576) = 0.99$:

$$P(\underbrace{3.312 - 2.576 * 0.341}_{2.434} < \mu_m - \mu_F < \underbrace{3.312 + 2.576 * 0.341}_{4.190}) = 0.99$$

statement <u>incorrect</u>.

(20) (b) Test the hypothesis that the variances are equal, using $\alpha = 5\%$.

$$\text{step 1)}\quad \frac{S_m'^2/\sigma_m^2}{S_F'^2/\sigma_F^2} \underset{\substack{\downarrow\\ H_0:\ \sigma_m^2=\sigma_F^2}}{=} \frac{S_m'^2}{S_F'^2} \sim F_{10,9}$$

step 2) Find quantiles:

$$q_{0.975}^{F_{10,9}} = 3.96\ ;\ q_{0.025}^{F_{10,9}} = \frac{1}{q_{0.975}^{F_{9,10}}} = \frac{1}{3.78} = 0.26$$

$$\text{step 3)}\quad F_{obs} = \frac{0.709}{0.518} = 1.37 \in (0.26, 3.96)$$

do <u>not</u> reject H_0

(10) (c) Which of the following statements is correct (only one statement is correct)?

○ If, for a given test, the significance level is equal to $\alpha\%$, then the power of that test is equal to $(1-\alpha)\,\%$.

○ The Type I error and the Type II error are equal for two-sided tests.

○ A pivotal quantity is a test statistic, if its repeated sampling distribution is completely known when the alternative hypothesis is true.

☒ None of the above.

6. A survey to 1000 portuguese individuals, regarding the outcome of COVID-19 infection, obtained the following results:

		Age		
		≤ 40	$(40,65)$	≥ 65
Outcome	Recovered	204	539	57
	Died	10	101	89

(20) (a) Based on an adequate test, at the 1% level, say if you agree with the following statement: “for those who died of COVID-19 infection, the distribution of age is discrete uniform (each age group has probability 1/3).“

$H_0 \cdot p_1 = p_2 = p_3 = 1/3 \qquad O_1 = 10,\ O_2 = 101,\ O_3 = 89$

$$E_1 = \frac{200}{3} = E_2 = E_3$$

$$Q_{obs} = \frac{(10-66.67)^2}{66.67} + \frac{(101-66.67)^2}{66.67} + \frac{(89-66.67)^2}{66.67} = 73.32 > 9.210$$

$q_{0.99}^{\chi^2_2} = 9.210 \Rightarrow$ <u>reject</u> H_0.

(20) (b) Is the outcome of COVID-19 infection independent of age? Use $\alpha = 5\%$.

$O_{11} = 204$	$O_{12} = 539$	$O_{13} = 57$	800
$O_{21} = 10$	$O_{22} = 101$	$O_{23} = 89$	200
214	640	146	1000

$E_{11} = \frac{800 * 214}{1000} = 171.2$	$E_{12} = 512$	$E_{13} = 116.8$
$E_{21} = 42.8$	$E_{22} = 128$	$E_{23} = 29.2$

$$Q \underset{H_0}{\sim} \chi^2_2 \text{ as } (r-1)(c-1) = 2 : q^{\chi^2_2}_{0.95} = 5.991$$

$$Q_{obs} = \frac{(204-171.2)^2}{171.2} + \dots + \frac{(89-29.2)^2}{29.2} = 191.6 > 5.991$$

we reject H_0

7. Consider the following models to explain the average monthly wage (in thousand euros) in terms of the number of years of education (estimated using OLS, standard errors in parentheses):

$$\ln\widehat{(wage_i)} = \underset{(0.4439)}{-1.9970} + \underset{(0.1756)}{1.2562}\ln(educ_i), \quad n = 428, \quad R^2 = 0.1073$$

$$\ln\widehat{(wage_i)} = \underset{(0.1852)}{-0.1852} + \underset{(0.0144)}{0.1086} educ_i, \quad n = 428, \quad R^2 = 0.1179$$

(15) (a) Interpret the estimated coefficients of the regressors $\ln(educ)$ and $educ$.

ln(educ): if I increase educ with 1%,
wage (on avg) increases with 1.2562%

(educ): if Increase educ with 1 year,
wage (on avg) increases with 10.86%

(15) (b) Test the significance of $\ln(educ)$ (i.e. test $H_0 : \beta_{\ln(educ)} = 0$).

$$\begin{cases} H_0 : \beta_{\ln(educ)} = 0 \\ H_1 : \beta_{\ln(educ)} \neq 0 \end{cases}$$

we did not assume that ε has a Normal distribution

$$T = \frac{\hat{\beta}_{\ln(educ)} - 0}{\sqrt{\hat{Var}(\hat{\beta}_{\ln(educ)})}} \overset{approx}{\sim} N(0,1) : z^{N(0,1)}_{0.975} = 1.960$$

$$T_{obs} = \frac{1.2562 - 0}{0.1756} = 7.15 > 1.96$$

we <u>reject</u> H_0

(10) (c) Based on your result in question b, can you say something about the confidence interval for $\beta_{\ln(educ)}$? Explain.

The 95% confidence interval for $\beta_{\ln(educ)}$ does <u>NOT</u> contain the value 0, because the test in Q7b rejected $H_0 : \beta_{\ln(educ)} = 0$.

You can use the space below and the other side of this page to add to your answers. Clearly state which question you are referring to.