30401 Mathematics and Statistics - Module 2 (Statistics) - BEMACS
GENERAL EXAM - - - 105 minutes

Surname		Name		Student Number	

I hereby confirm my attendance at the exam.
I declare I have read the Exam rules and I commit to respect them.
Signature:

Some exercises refer to the dataframe **Sleep** in the file **Sleep.Rdata**. The dataframe contains information on sleep habits, demographics, physical activity, stress levels, and symptoms of sleep disorders, collected on a sample of subjects with specific occupations. Explanations of the meaning of the variables are provided in the text when needed.

Exercise 1 (5 points)

Now let X and Y be two independent random variables such that $X \sim$ Be(p) is a Bernoulli r.v. with parameter p and $Y \sim Exp(1)$ is an exponential distribution with parameter 1.

a) (2pt*) For each of the following statements decide whether it is **TRUE** or **FALSE** in general.

- $E[XY] = p$ [TRUE] [FALSE]
- $P(XY = 0) = 0$ [TRUE] [FALSE]

b) (3pt**) Calculate Var$[XY]$. Carefully report the proceedings.

Exercise 2 (16 points, R Dataset)

Empirical studies on sleep patterns suggest that the nightly sleep duration X (in hours) of an individual can be modelled according to the probability density function $P(x;k,\lambda)$ defined

$$P(x;k,\lambda)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$$

for $x \geq 0, k \geq 2, \lambda > 0$. Throughout questions a)-e) take the parameter λ to be fixed at $\lambda = 8$, and treat the parameter k as the only unknown parameter that we want to **learn** from sample data.

a) (3pt*) Draw a qualitative sketch of the pdf $P(x;k,\lambda)$ for $\lambda = 8$ and $k = 5$. Briefly explain how increasing the value of k (with fixed λ) affects the mode and the concentration of the distribution.

Let $x_1, \ldots, x_n$ be an independent sample of sleep–duration observations.

b) (3pt*) Write explicitly the likelihood function $l(k)$ and the negative log-likelihood $L(k)$ of the parameter k given the generic sample data.

c) (3pt*) Compute the derivative of $L(k)$ with respect to k. State the condition that must be satisfied by the maximum-likelihood estimator $\hat{k}$.
(Hint: use the following result to differentiate L with respect to k.

$$\frac{d}{dk}\left(\frac{x_i}{\lambda}\right)^k=\left(\frac{x_i}{\lambda}\right)^k\cdot\log\left(\frac{x_i}{\lambda}\right))$$

As the condition in c) cannot be solve analytically, you proceed to find the maximum likelihood estimate numerically.

d) (3pt**) Using the sample contained in the variable **SleepDuration**, compute and report the value of $L(k)$ for $k = 8,9,10$. Which of these values of k gives the best fit according to the negative log-likelihood criterion?

e) (2pt**) Proceed and obtain numerically the maximum likelihood estimate for k to two decimal places.

f) (2pt***) It is well known that the expectation of the distribution X is:

$$E[X]=\lambda\cdot\Gamma\left(1+\frac{1}{k}\right)$$

where $\Gamma(\cdot)$ is the gamma function. You can evaluate in R the gamma function for any value with the command `gamma()`.
Assume now **both parameters** λ and k are **unknown**. Compute the sample mean $\bar{x}$ of the variable **SleepDuration**.

Calculate and report the maximum likelihood estimates $\hat{k}$ and $\hat{\lambda}$ of k and λ respectively, subject to the following constraint:

$$\hat{\lambda}\cdot\Gamma\left(1+\frac{1}{\hat{k}}\right)=\bar{x}.$$

(Hint: minimise the negative log-likelihood numerically. For every possible value of k, the value of λ can be determined by the above constraint)

A

Exercise 3 (10 points, R Dataset)

a) (5pt*) We are interested in verifying the null hypothesis that in the population the proportion of individuals who suffer from any sleep disorders (**SleepDisorder** = Insomnia or Other) is not higher than 0.35 against the alternative hypothesis that it is higher than 0.35. Considering the sample data, answer the question through a suitable hypothesis test. Specify:

   i) (*) the null and alternative hypothesis

   ii) (*) detailed derivation of the p-value

   iii) (*) a rigorous definition of the p-value

   iv) (*) your final conclusion

b) (5pt**) We are interested in the average duration of sleep (variable **SleepDuration**) and in the possible differences across subjects with different occupations (variable **Occupation**). In particular, do we have enough statistical evidence to state that the average duration of sleep among nurses (**Occupation** = Nurse) is higher than that among doctors (**Occupation** = Doctor)? Construct a suitable hypothesis test to answer. Specify:

   i) the null and alternative hypothesis

   ii) detailed derivation of the p-value

   iii) your final conclusion

A