B

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 (6 points, R Dataset)

a) (2pt*) Briefly describe the distribution of the quality of sleep (variable **SleepQuality**) in terms of central tendency, dispersion and shape. Report the values of all the relevant quantities you rely on for your answer. Also report the name of a suitable graphical representation for this variable.

b) (0.5pt*) What is the threshold separating the 10% of the subjects with the best quality of sleep (variable **SleepQuality**) from the others?

- ☐ 4.82
- ☐ 5.44
- ☐ 9.26
- ☐ 9.67

c) (1.5pt*) Below which threshold is a subject considered a lower outlier with respect to quality of sleep? Report such threshold and the number and values of such lower outliers in the dataset.

d) (1pt*) Is there an association between quality of sleep (variable **SleepQuality**) and **Age**? Report both the summary measure and graphical representation suitable to answer and comment clearly on the type, direction and strength of the association.

e) (1pt*) Is the sample percentage of subjects who suffer from some sleep disorder (**SleepDisorder** = Insomnia or Other) higher among those with Normal **BMI** or with Overweight **BMI**? Report the percentages you rely on to answer.

EXERCISE 2 (4 points)

A financial investment will produce a profit equal to W. Since the investment involves some risk, the profit W is assumed to be a normally distributed random variable with mean equal to 250 ('000 euros) and standard deviation equal to 100 ('000 euros).

a) (0.5pt*) What is the probability that the investment will result in a financial loss (i.e. that the profit is negative)?

- ☐ 0.62%
- ☐ 0%
- ☐ 34.46%
- ☐ 49%

b) (1pt**) What is the minimum loss generated from the investment in the worst 0.1% scenarios? (This is called the 0.1*% Value at Risk* of the investment).

c) (2.5pt**) Assume there are other independent investments with similar profitability and risk profile. Call W_i the profit of the $i-th$ investment and assume $W_1, W_2, \ldots, W_{100}$ are i.i.d. random variables such that $W_i \sim \mathcal{N}(250,100^2)$.

i) What is the approximate probability that 3 or more out of the 100 investments result in a financial loss?

ii) What is the probability that the average profit of the 100 investments is negative?

B

Exercise 3 (4 points)

Part I

Consider a continuous random variable T with unknown distribution, that represents a positive waiting time before an individual falls asleep. Assume that $E[T] = 20$ and $Var(T) = 100$.
(Hint: use Chebyshev's)

a) (1pt*) Provide a suitable bound for the probability that such waiting time is above 60.

b) (1pt**) Find the minimum c such that the following inequality always holds: $P(T \geq c) \leq 0.01$.

Part II

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.

c) (1pt*) For each of the following statements decide whether it is **TRUE** or **FALSE** in general.

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

d) (1pt**) Calculate Var$[XY]$. Carefully report the proceedings.

Exercise 4 (11 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) (2pt*) Draw a qualitative sketch of the pdf $P(x;k,\lambda)$ for $\lambda = 8$ and $k = 10$. 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) (2pt*) Write explicitly the likelihood function $l(k)$ and the negative log-likelihood $L(k)$ of the parameter k given the generic sample data.

c) (2pt*) 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) (2pt**) Using the sample contained in the variable **SleepDuration**, compute and report the value of $L(k)$ for $k = 8.5, 9.5, 10.5$. Which of these values of k gives the best fit according to the negative log-likelihood criterion?

e) (1pt**) 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)

B

Exercise 5 (6 points, R Dataset)

a) (3pt*) 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 0.40 against the alternative hypothesis that it is higher than 0.40. 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) (3pt**) 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 doctors (**Occupation** = Doctor) is lower than that among nurses (**Occupation** = Nurse)? 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