30401 Mathematics and Statistics - Module 2 (Statistics) - BEMACS
FIRST PARTIAL EXAM, Version A - 60 minutes

Surname		Name		Student Number	

I commit to respect the Bocconi Honour Code. Signature:

For the following questions, refer to the dataset **Employee**, contained in the file **Employee.Rdata**. To assess employee satisfaction, productivity, and fair compensation, the consulting company **Bill&Able Ltd**. conducts a thorough examination of a sample of its workforce. The dataset contains various variables describing a sample of employees at **Bill&Able**. Explanations of the variables meaning are provided in the text when needed.

Exercise 1 (10 points, R Dataset)
With reference to **Bill&Able**, we are interested in the relationship between employees' salaries (variable **Salary**) and their department (variable **Department**)

a) (2pt*) Propose a suitable plot to highlight the differences in the distribution of salaries across departments. Sketch the proposed plot and, based solely on it, determine whether there is a relation between the employees' salary and their department.

b) (2pt*) Define rigorously and in full generality the mean and the median and explain at least one advantage of each of them over the other for comparing distributions.

c) (2pt*) Calculate and report the mean salary of employees in each department and write your considerations.

d) (2pt**) What is the range of salaries for the top 15% earners in the HR department (**Department**=HR)?

e) (2pt**) What is the department with the highest proportion of senior employees (**Role**=Senior)? Motivate your answer and report all the measures you rely on.

EXERCISE 2 (12 points, R Dataset and Frequency Table)

The company **Bill&Able** wants to compare the salary distribution of its employees (see the **Salary** variable in the **Employee** dataset) with that of the employees of a competing firm: **Con$ult Ltd.**

From a published report, the salary distribution of a sample of 200 employees of **Con$ult** is provided below, categorised into interval classes.

Salary Class	(1000,1500]	(1500,2000]	(2000,3000]	(3000,5000)	(5000,8000]
Employees of Con$ult	90	40	10	10	50

a) (1pt*) Calculate and report the median, the first quartile and third quartile of the salaries for the sample of employees at **Bill&Able** (dataset).

b) (1pt*) For each Salary Class in the frequency table of the salaries of employees at **Con$ult** (table), report its relative frequency and frequency density. (*You can use the rows in the table*)

c) (2pt*) Calculate and report the median, the first quartile and third quartile of the salaries for the sample of employees at **Con$ult** (table).

d) (2pt*) For each of the two samples (table and dataset) calculate and report the means and standard deviations of the salaries.

e) (2pt**) Using the calculated quartiles, medians, means and standard deviations compare the two distributions of salaries for employees of the two companies. In particular comment, briefly but with rigorous justifications, the following aspects:

   i) Centrality

   ii) Dispersion

   iii) Shape

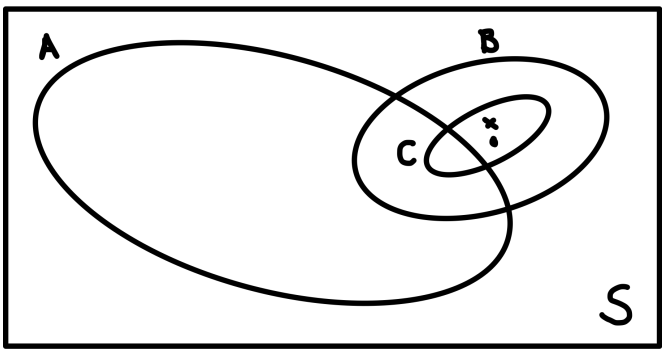
   iv) Disparity between top and bottom earners

f) (2pt**) Consider the sample distribution of Salaries of employees of **Bill&Able** (dataset). Calculate and report the threshold above which a salary is considered an upper outlier. Also, report how many employees in the sample have salaries above such threshold, their salaries and the department they work for.

g) (2pt**) Consider the threshold for upper outliers calculated at previous point (if you could not calculate the threshold, assume 6500 as the threshold). What is the percentage of employees of **Con$ult** (table) in the sample with a salary above this threshold?
What is the percentage of employees of **Con$ult** (table) in the sample earning above 5000 instead?
Explain whether these figures are exact or approximated and justify your answer.

EXERCISE 3 (3 points)

(3pt*) Let A, B and C be 3 events as represented in the Venn diagram below, and let x be an element belonging to event C.



For each of the following statements decide whether it is TRUE or FALSE in general.

- $P(\{x\}) \geq P(C)$ **[TRUE] [FALSE]**
- $P(B \cup C) = P(C)$ **[TRUE] [FALSE]**
- $P(A \cup B \cup C) = P(A) + P(B) - P(A \cap B)$ **[TRUE] [FALSE]**
- $(A \cup B)^c = \varnothing$ **[TRUE] [FALSE]**
- $x \in B \cap C^c$ **[TRUE] [FALSE]**
- $C \in x$ **[TRUE] [FALSE]**

EXERCISE 4 (6 points)

The consulting company **Synergeeks Ltd.** allows its employees to work from home occasionally but aims to keep remote work to a minimum. To monitor this, the company tracks the number of employees who choose to work from home each day at one of its offices. On a typical day, this number X appears to follow a Poisson distribution with a mean of 20 employees, i.e. X~Poisson(μ=20).

a) (1pt*) According to the Poisson model, roughly how many employees decide to work from home on the 90% most typical days?

b) (1pt*) If 30 or more employees work from home on a given day, it may cause operational challenges for the company. What is the probability of this occurring?

c) (2pt**) Over a year with 200 working days, on how many days can the company expect, on average, to experience the disruption of 30 or more employees working from home? Additionally, what is the probability that this disruption occurs at most 10 times in a year?

d) (2pt***) (*For this question report all the R code used to produce your simulations*)
The company revises its model for the number of employees working from home, as it appears to be strongly influenced by weather conditions. Specifically:

- On rainy days, the number of employees X_1 working from home follows a Poisson distribution with a mean of 30: $X_1 \sim \text{Poisson}(\mu_1 = 30)$.
- On non-rainy days, the number of employees X_2 working from home also follows a Poisson distribution, but with a lower and unknown mean: $X_2 \sim \text{Poisson}(\mu_2)$ for some $\mu_2 < \mu_1$.
- Across all days (both rainy and non-rainy), the average number of employees working from home is 20, and on any day the probability that it rains is 20%.

Using this revised and more realistic model, update your answers to the previous questions either through simulation or analytically (though the latter may be more challenging). Specifically, compute:

i) The parameter μ_2

ii) The number of employees working from home on the 90% most typical days

iii) The probability that 30 or more employees work from home on a given day.

iv) The probability that this disruption (30 or more employees working from home) occurs at most 10 times out of 200 working days.