Ministerie van Defensie

Department of Military Sciences

Student data	
Name:	
Peoplesoft number:	
Class:	
Signature:	

General			
Course:	Probability & Statistics (Resit)	**Course code:**	P&S
Date:	4 juli 2025	**Time:**	10:00-13:00
Examiner:	Dr. ir. D.A.M.P. Blom	**Number of Pages:**	6
Number of Questions	5	**Total Points:**	80

General instructions

- All answers must be supported by a clear explanation. Answers such as “yes” or “no” without justification will receive no credit.
- Round final answers to four decimal places, where applicable.
- If you are unable to solve a subquestion, you are encouraged to make a reasonable assumption and proceed. Partial credit may still be awarded for correct methodology, even if intermediate answers are incorrect.
- The use of a graphical calculator without a CAS (Computer Algebra System) is permitted.
- No exam-related material may be taken out of the examination room.
- Please write your name and PeopleSoft number on each page and number all pages of your answers (e.g., 1/5, 2/5, etc. if you hand in five answer sheets).
- The use of electronic devices capable of sending, receiving or storing information (e.g., mobile phones, smartwatches) is strictly prohibited. These must be left outside the exam room or handed in to the examiner, switched off or in airplane mode.
- Ensure your handwriting is legible. Illegible or unclear answers will not be graded.
- Toilet visits are only allowed with prior permission from the examiner.
- Upon leaving the examination room, all materials (exam paper, scrap paper, formula sheets) must be handed in to the examiner.

Grading
- The final grade of the Probability and Statistics course is entirely based on this exam. - The exam consists of five open-ended questions, each with subquestions. - The number of points available for each (sub)question is indicated in brackets. A total of 90 points can be earned. - Your final grade will be calculated by dividing the total points earned by 9. - A minimum final grade of 5.5 is required to pass the course.

Procedure after the exam
- Exam results will be published within ten working dates after the exam date. - If you have questions about the grading, you may contact the course coordinator within ten working days after the results have been released.

Good luck!

Opgave 1 (15 punten) A set of n independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ are drawn from a distribution with the probability density function:

$$f(x;\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, \quad x \geq 0, \quad \theta > 0$$

1a [3pt] Write down the likelihood function $L(x_1, x_2, \ldots, x_n; \theta)$ given a sample of realizations of $X_1, X_2, \ldots, X_n$.

Uitwerking

The likelihood function can be written as follows:

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta) &= f(x_1;\theta) \cdot f(x_2;\theta) \ldots \cdot f(x_n;\theta) \\ &= \frac{1}{\theta}e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta}e^{-\frac{x_2}{\theta}} \ldots \cdot \frac{1}{\theta}e^{-\frac{x_n}{\theta}} \\ &= (\frac{1}{\theta})^n e^{-\frac{\sum_{i=1}^n x_i}{\theta}} \end{aligned}$$

1b [7pt] Derive the maximum likelihood estimator (MLE) of the parameter θ.

Hint: consider the natural logarithm (ln) of the likelihood function (the log-likelihood).

Uitwerking

The maximum likelihood estimator can be obtained by first considering the log-likelihood, as it results in easier computations later on:

$$\begin{aligned} \ell(x_1, x_2, \ldots, x_n; \theta) &= \ln{(\frac{1}{\theta})^n e^{-\frac{\sum_{i=1}^n x_i}{\theta}}} \\ &= \ln{(\frac{1}{\theta})^n} + \ln{e^{-\frac{\sum_{i=1}^n x_i}{\theta}}} \\ &= -n \cdot \ln\theta - \frac{\sum_{i=1}^n x_i}{\theta} \end{aligned} \qquad \text{(3pt)}$$

The maximum likelihood estimator (MLE) can then be found by optimizing the log-likelihood function, i.e., taking the derivative of the likelihood function and setting it equal to zero: Notice that the log-likelihood function has the same optimal solutions as the likelihood function, as the logarithm is a strictly increasing function and hence preserving optimality.

$$\frac{d\ell(x_1, x_2, \dots, x_n; \theta)}{d\theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0 \qquad \text{(1pt)}$$

Solving the equation gives us:

$$-\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0 \Rightarrow \frac{\sum_{i=1}^{n} x_i}{\theta^2} = \frac{n}{\theta} \Rightarrow \hat{\theta}_{\text{MLE}} = \frac{\sum x_i}{n} = \overline{x}. \qquad \text{(2pt)}$$

The maximum likelihood estimator of θ is the sample mean $\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$. (1pt)

1c [5pt] Show that the sample mean $\overline{X} = \frac{(X_1+X_2+\ldots+X_n)}{n}$ is an unbiased estimator for the parameter θ.

Uitwerking

We need to show that $E[\frac{\sum_{i=1}^{n} X_i}{n}] = \theta$. Notice that the given probability density function is equal to the density function of a exponential distribution with parameter $\lambda = \frac{1}{\theta}$. (2pt)

This means that each X_i has an expected value of $E[X_i] = \frac{1}{\lambda} = \theta$. (1pt)

Therefore, we have that

$$E[\frac{\sum_{i=1}^{n} X_i}{n}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(\theta + \theta + \ldots + \theta) = \frac{1}{n} \cdot n\theta = \theta,$$

which finishes the proof. (2pt)

Opgave 2 (20 punten) A coastal rescue team is conducting search operations for missing swimmers lost at sea. Based on previous data and ocean current modeling, the joint probability density function (PDF) for the coordinates (X, Y) of a swimmer's location (in kilometers relative to the shore) is given by:

$$f_{X,Y}(x,y) = \begin{cases} C \cdot (3x+2y), & \text{if } 0 \leq x \leq 2 \text{ and } 0 \leq x \leq 3 \\ 0, & \text{otherwise,} \end{cases}$$

where C is a constant.

2a [6pt] Determine the value of C so that $f_{X,Y}(x,y)$ satisfies the properties of a valid joint probability density function.

Uitwerking

The joint probability density function $f_{X,Y}(x,y)$ is a valid joint probability density function if the following two conditions hold

1. First condition: $f_{X,Y}(x,y) \geq 0$ for all possible values of x and y, and (1pt)
2. Second condition: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy = 1$. (1pt)

The first condition holds if and only if $C(3x+2y)$ for $0 \leq x \leq 2$ and $0 \leq y \leq 3$, so we need $C \geq 0$. The second condition holds if

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy \\ &= \int_0^3 \int_0^2 C \cdot (3x+2y)\ dx\ dy \\ &= C \cdot \int_0^3 \left[\int_0^2 (3x+2y) dx \right]\ dy \\ &= C \cdot \int_0^3 \left[\frac{3}{2}x^2 + 2xy \right]_0^2\ dy \\ &= C \cdot \int_0^3 6 + 4y\ dy \\ &= C \cdot \left[6y + 2y^2 \right]_0^3 \\ &= C \cdot (18+18) = 36 \cdot C \end{aligned}$$ (3pt)

The constant C should be equal to $\frac{1}{36}$ in order for $f_{X,Y}(x,y)$ to satisfy the properties of a valid joint probability density function.

2b [6pt] Compute the marginal PDFs $f_X(x)$ and $f_Y(y)$, for X and Y respectively.

Uitwerking

– The marginal PDF of X is given by

$$\begin{aligned} f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}\ dy &= \int_0^3 \frac{1}{36} \cdot (3x + 2y)\ dy \\ &= \left[\frac{1}{12}xy + \frac{1}{36}y^2\right]_0^3 \\ &= \frac{1}{4}x + \frac{1}{4} \end{aligned}$$ (3pt)

– The marginal PDF of Y is given by

$$\begin{aligned} f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}\ dx &= \int_0^2 \frac{1}{36} \cdot (3x + 2y)\ dx \\ &= \left[\frac{1}{24}x^2 + \frac{1}{18}xy\right]_0^2 \\ &= \frac{1}{9}y + \frac{1}{6} \end{aligned}$$ (3pt)

2c [8pt] Are X and Y independent random variables? Justify your answer with a computation.

Uitwerking

To show whether or not X and Y are independent random variables, we need to compute their covariance and check if it is equal to zero (independence) or not (not independence). We can compute the covariance as follows:

$$Cov(X, Y) = E[XY] - E[X]E[Y].$$ (1pt)

We first need to compute these quantities:

$$\begin{aligned} E[X] = \int_{-\infty}^{\infty} x f_X(x)\ dx &= \int_0^2 x \cdot \left(\frac{1}{4}x + \frac{1}{4}\right)\ dx \\ &= \int_0^2 \frac{1}{4}x^2 + \frac{1}{4}x\ dx \\ &= \left[\frac{1}{12}x^3 + \frac{1}{8}x^2\right]_0^2 = \frac{7}{6}. \end{aligned}$$ (2pt)

$$
\begin{aligned}
E[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy &= \int_0^3 y \cdot \left(\frac{1}{9}y + \frac{1}{6}\right)\, dy \\
&= \int_0^3 \frac{1}{9}y^2 + \frac{1}{6}y\, dy \\
&= \left[\frac{1}{27}y^3 + \frac{1}{12}y^2\right]_0^3 = \frac{7}{4}. && \text{(1pt)}
\end{aligned}
$$

$$
\begin{aligned}
E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\, dx\, dy &= \int_0^3 \int_0^2 \left(\frac{1}{12}x^2y + \frac{1}{18}xy^2\right)\, dx\, dy \\
&= \int_0^3 \left[\frac{1}{36}x^3y + \frac{1}{36}x^2y^2\right]_0^2\, dy \\
&= \int_0^3 \frac{2}{9}y + \frac{1}{9}y^2\, dy \\
&= \left[\frac{1}{9}y^2 + \frac{1}{27}y^3\right]_0^3 = 2. && \text{(2pt)}
\end{aligned}
$$

The covariance of X and Y is equal to $\text{Cov}(X,Y) = E[XY] - E[X]\cdot E[Y] = 2 - \frac{7}{6}\cdot\frac{7}{4} = -\frac{1}{24} \neq 0$. Hence, X and Y cannot be independent random variables. (1pt)

Opgave 3 (22 punten) In a joint project between physical training instructors of the Royal Military Academy (KMA) and the medical service, the relationship is investigated between average sleep duration of a cadet (in hours) and the recovery time after intense physical exertion (in hours).

Recovery time is measured using a muscle soreness index; a cadet is considered fully recovered when this index falls below a certain threshold.

Voor de hersteltijd wordt gekeken naar een spierpijn-index, en een cadet is volledig hersteld wanneer de spierpijn-index onder een gegeven drempelwaarde komt.

For nine cadets, the average sleep duration and recovery time after intense exertion are measured.

Sleep duration (in hours)	$4,0$	$4,5$	$5,0$	$5,5$	$6,0$	$6,5$	$7,0$	$7,5$	$8,0$
Recovery time (in hours)	66	61	63	62	65	64	57	59	60

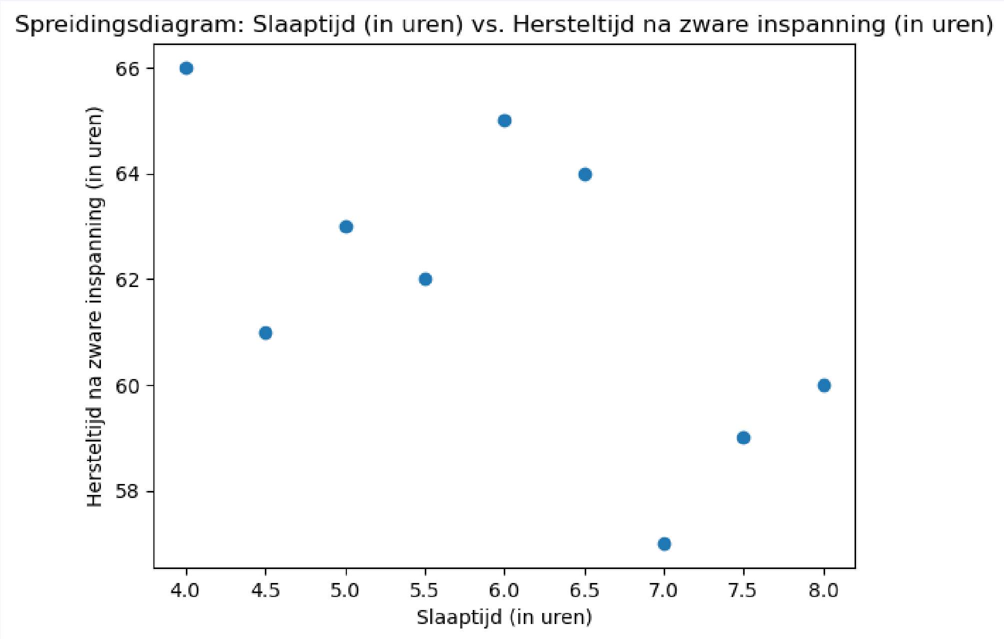
3a [2pt] If we want to perform a regression analysis, which variable would be the dependent variable Y and which variable the independent variable X?

Uitwerking

In regression analysis, it makes the most sense to take sleep time as the independent variable X and recovery time as the dependent variable Y. A longer recovery time does not explain why someone sleeps longer, but the other way around might make sense. (2pt)

3b [5pt] Draw the corresponding scatter plot based on your answer to subquestion (a).

Uitwerking



3c [8pt] Calculate Pearson's correlation coefficient $r(x, y)$. What can you conclude about the relationship between the two variables?

Uitwerking

We start by calculating Pearson's correlation coefficient:

$$r(x,y) = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2) \cdot (\overline{y^2} - \overline{y}^2)}}$$

We use the following table to find the necessary values:

x	y	xy	x^2	y^2
4	66	264	16	4356
$4,5$	61	$274,5$	$20,25$	3721
5	63	315	25	3969
$5,5$	62	341	$30,25$	3844
6	65	390	36	4225
$6,5$	64	416	$42,25$	4096
7	57	399	49	3249
$7,5$	59	$442,5$	$56,25$	3481
8	60	480	64	3600
$\overline{x} = 6$	$\overline{y} = 61,8889$	$\overline{xy} = 369,1111$	$\overline{x^2} = 37,6667$	$\overline{y^2} = 3837,8889$

(4pt)

Now we compute Pearson's correlation coefficient as follows:

$$\begin{aligned} r(x,y) &= \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x}^2 - \overline{x^2}) \cdot (\overline{y}^2 - \overline{y^2})}} \\ &= \frac{369,1111 - 6 \cdot 61,8889}{\sqrt{(6^2 - 37,6667) \cdot (61,8889^2 - 3837,8889)}} \\ &= \frac{-2,2222}{3,5717} \\ &\approx -0,6222. \end{aligned}$$

(3pt)

The correlation coefficient is clearly negative and not close to zero, indicating a fairly strong negative trend. However, there is still some uncertainty in the exact linear relationship. (1pt)

3d [7pt] Calculate the regression line $Y = a + b \cdot X$ by computing the coefficients a and b. Based on this regression line, give a statistically sound prediction for the recovery time of a cadet who has slept for 6 hours and 45 minutes.

Uitwerking

In order to compute the coefficients a and b, we reuse the table from subquestion (a). We have that

$$\begin{aligned} b &= \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \\ &= \frac{369,1111 - 6 \cdot 61,8889}{37,6667 - (6)^2} \\ &= \frac{-2,2222}{1,6667} \approx -1,3333 \\ a &= \overline{y} - b \cdot \overline{x} && \text{(2pt)} \\ &= 61,8889 - (-1,3333 \cdot 6) \\ &\approx 69,8889. && \text{(2pt)} \end{aligned}$$

The formula of the regression line is equal to $Y = 69,8889 - 1,3333 \cdot X$. A statistically sound prediction of the recovery time of a cadet who slept for 6 hours and 45 minutes is found by filling in $X = 6,75$: This results in a predicted recovery time of $Y = 69,8889 - 1,3333 \cdot 6,75 = 60,88887$ hours, or just under 61 hours. (1pt) (2pt)