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!

Problem 1 (17 points) During routine patrols in the North Sea, a helicopter unit detects multiple Russian vessels suspected of attempting to damage undersea communication cables. The helicopter record two discrete variables hourly:

- X: the number of radar contacts detected per hour (values: $1, 2, 3$)
- Y: the number of confirmed hostile aircraft identified per hour (values: $1, 2, 3$)

The joint probability distribution of X and Y reflects the likelihood of these events occurring simultaneously during patrol hours.

X Y	1	2	3
1	0.20	0.12	0.08
2	0.05	0.18	0.10
3	0.07	0.03	0.17

1a [6pt] Calculate the marginal distributions of X and Y, as well as $E[X]$, Var(X), $E[Y]$ and Var(Y).

Solution

The marginal distribution functions f_X and f_Y can be determined by calculating row totals and columns totals respectively. This gives:

$X\backslash Y$	1	2	3	$\boldsymbol{f_X}$
1	0.20	0.12	0.08	**0.40**
2	0.05	0.18	0.10	**0.33**
3	0.07	0.03	0.17	**0.27**
$\boldsymbol{f_Y}$	**0.32**	**0.33**	**0.35**	**1**

(2pt)

The requested expected values can be calculated using

$$E[X] = 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3)$$

$$= 1 \cdot 0.40 + 2 \cdot 0.33 + 3 \cdot 0.27 = 1.87$$

$$E[Y] = 1 \cdot P(Y = 1) + 2 \cdot P(Y = 2) + 3 \cdot P(Y = 3) =$$

$$= 1 \cdot 0.32 + 2 \cdot 0.33 + 3 \cdot 0.35 = 2.03$$

(2pt)

The requested variances can be calculated using

$$\begin{aligned}\text{Var}(X) &= (1-1.87)^2 \cdot P(X=1) + (2-1.87)^2 \cdot P(X=2) + (3-1.87)^2 \cdot P(X=3) \\ &= 0.6531 \\ \text{Var}(Y) &= (1-2.03)^2 \cdot P(Y=1) + (2-2.03)^2 \cdot P(Y=2) + (3-2.03)^2 \cdot P(Y=3) \\ &= 0.6691\end{aligned}$$

(2pt)

1b [4pt] Calculate the covariance $\text{Cov}(X,Y)$.

Solution

The covariance can be calculated using the formulas

$$\text{Cov}(X,Y) = E[XY] - E[X] \cdot E[Y]$$

(1pt)

Therefore, we need to compute $E[XY]$, which is equal to

$$E[XY] = \sum_{x=1}^{3} \sum_{y=1}^{3} x \cdot y \cdot P(X=x, Y=y) = 1 \cdot 1 \cdot 0.20 + 1 \cdot 2 \cdot 0.12 + \ldots + 3 \cdot 3 \cdot 0.17 = 4.02.$$

(2pt)

The covariance is therefore equal to

$$\text{Cov}(X,Y) = E[XY] - E[X] \cdot E[Y] = 4.02 - 1.87 \cdot 2.03 = 0.2239$$

(1pt)

1c [5pt] Calculate the following probabilities:

- $P(X=3, Y \leq 2)$
- $P(X=3 \mid Y \leq 1)$
- $P(Y \leq 1 \mid X \leq 2)$

Solution

For the first probability, we get:

$$P(X=3, Y \leq 2) = P(X=3, Y=1) + P(X=3, Y=2) = 0.07 + 0.03 = 0.10$$

(1pt)

For the second probability, we get:

$$\begin{aligned}
P(X \geq 2 \mid Y = 2) &= \frac{P(X \geq 2, Y = 2)}{P(Y = 2)} \\
&= \frac{P(X = 2, Y = 2) + P(X = 3, Y = 2)}{P(Y = 2)} \\
&= \frac{0.18 + 0.03}{0.33} \\
&\approx 0.6364
\end{aligned}$$

(2pt)

For the third probability, we get:

$$\begin{aligned}
P(Y = 3 \mid X \geq 2) &= \frac{P(Y = 3, X \geq 2)}{P(X \geq 2)} \\
&= \frac{P(Y = 3, X = 2) + P(Y = 3, X = 3)}{P(X = 2) + P(X = 3)} \\
&= \frac{0.10 + 0.17}{0.33 + 0.27} \\
&= 0.45
\end{aligned}$$

(2pt)

1d [2pt] Are X and Y independent? Explain your answer.

Solution

If X and Y are independent random variables, then the covariance Cov$(X, Y) = 0$. As the covariance is non-zero in this case, X and Y must be dependent random variables.

(1pt)

(1pt)

Problem 2 (15 points) In a military exercise, measurements are conducted regarding the time a certain communication signal needs to reach a control post. These times $X_1, X_2, \dots, X_n$ are assumed to be independent and identically distributed according to a uniform distribution over the interval $[\theta; \theta+1]$, where θ is an unknown real parameter.

2a [5pt] Write down the likelihood function $L(x_1, x_2, \dots, x_n; \theta)$ for the parameter θ in terms of a sample of realizations $x_1, x_2, \dots, x_n$.

Solution

The probability density function for a single observation x is

$$f(x;\theta) = \begin{cases} 1, & \text{if } \theta \le x \le \theta + 1 \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(1pt)}$$

The likelihood function is therefore the product of ones and zeros, depending on the sample of observations $x_1, x_2, \dots, x_n$. In particular, we get

$$\begin{aligned} L(x_1, x_2, \dots, x_n; \theta) &= f(x_1;\theta) \cdot f(x_2;\theta) \dots \cdot f(x_n;\theta) \\ &= 1(\theta \le x_1 \le \theta+1) \cdot 1(\theta \le x_2 \le \theta+1) \cdot \dots \cdot 1(\theta \le x_n \le \theta+1) \end{aligned} \qquad \text{(2pt)}$$

This means we can write the likelihood function as follows:

$$L(x_1, x_2, \dots, x_n; \theta) = \begin{cases} 1, & \text{if } \theta \le x_i \le \theta + 1 \text{ for all } i = 1, \dots, n \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(2pt)}$$

2b [5pt] Derive the maximum likelihood estimator (MLE) of the parameter θ.

Solution

In order for the likelihood function $L(x_1, x_2, \dots, x_n; \theta)$ to take on the maximum value 1, every observation x_i needs to lie between $\theta \le x_i \le \theta + 1$, hence we need that $\theta \le \min_i x_i$ and $\theta + 1 \ge \max_i x_i$. This means the likelihood is constant as long (2pt) as θ is in the interval $[\max_i x_i - 1; \min_i x_i]$. (1pt)

Since the likelihood is constant on this interval, any θ in this range is a maximum likelihood estimator. Commonly, the right endpoint is chosen, that is $\hat{\theta}_{MLE} = \min_i x_i$. (2pt)

2c [5pt] Show that the estimator $\overline{X} - \frac{1}{2} = \frac{(X_1+X_2+...+X_n)}{n} - \frac{1}{2}$ is an unbiased estimator for the parameter θ.

Solution

If $\hat{\theta} = \overline{X} - \frac{1}{2}$ is an unbiased estimator, it should hold that $E[\overline{X} - \frac{1}{2}] = \theta$. (1pt)

Notice that the expected value of a single measurement X_i is equal to

$$E[X_i] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_{\theta}^{\theta+1} x \cdot 1\, dx = \theta + \frac{1}{2}.$$ (1pt)

Therefore, the expected value of the sample mean $\overline{X}$ is also equal to $\theta + \frac{1}{2}$. Rearranging the terms gives $E[\overline{X}] - \frac{1}{2} = E[\overline{X} - \frac{1}{2}] = \theta$. This shows that $\overline{X} - \frac{1}{2}$ is an unbiased estimator. (1pt) (1pt) (1pt)

Problem 3 (20 points) A naval research unit investigates the signal strength (in decibels, dB) of sonar pulses reflected from a newly designed stealth submarine hull.

Under standard conditions, the mean reflected signal strength from a conventional hull is $\mu_0 = 65$ dB, with unknown standard deviation σ dB.

To assess whether the new stealth design reduces detectability, a test series is conducted. Signal strength is measured in 10 independent trials with the following results (in dB):

$$\{63.2,\ 64.1,\ 62.5,\ 63.8,\ 65.0,\ 64.7,\ 63.5,\ 62.9,\ 63.6,\ 64.2\}$$

Assume the measurements follow a normal distribution. Use a significance level of $\alpha = 0.05$.

3a [4pt] Determine the mean and standard deviation of this sample.

Solution

We can compute the sample mean and the sample variance as follows:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{63.2 + 64.1 + \ldots + 64.2}{10} = \frac{637.5}{10} = 63.75 \qquad \text{(1pt)}$$

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \frac{1}{10-1} \cdot \left((63.2 - 63.75)^2 + (64.1 - 63.75)^2 + \ldots + (64.2 - 63.75)^2 \right) \\ &\approx 0.6072 \end{aligned} \qquad \text{(2pt)}$$

The sample standard deviation is then obtained by taking the square root of the sample variance:

$$s = \sqrt{0.62} \approx 0.7792 \qquad \text{(1pt)}$$

3b [3pt] State the null and alternative hypotheses, and explain the direction of the test.

Solution

Since we aim at testing whether the mean reflected signal strength is still $\mu_0 = 65$ or if it has decreased, the null hypothesis H_0 and the alternative hypothesis H_1 of this test can be formulated as follows:

$$H_0 : \mu = 65 \quad \text{(no change in reflectivity)} \qquad \text{(1pt)}$$

$$H_1 : \mu < 65 \quad \text{(stealth hull reduces reflectivity)} \qquad \text{(1pt)}$$

This is a **one-sided** left-tailed test. (1pt)

3c [9pt] Perform the hypothesis test for the mean reflected signal strength μ based on the critical region.

Solution

Since the sample size $n = 10$ is smaller than 30, and the population standard deviation σ is unknown, we need to resort to the t-distribution. This t-distribution (1pt) has $\text{df} = n - 1 = 9$ degrees of freedom. (1pt)

We want to compute a critical region for the population mean μ with significance level $\alpha = 0.05$. The t-value is equal to

$$t = \text{InvT}(\text{area} = 1 - \alpha; \text{df} = n - 1) = \text{InvT}(\text{area} = 0.95; \text{df} = 9) = 1.8331. \qquad \text{(2pt)}$$

Since the hypothesis is left-tailed, the critical region is of the form $(-\infty, g]$. In (1pt) particular, we can compute the boundary g as follows

$$\begin{aligned} g &= \mu - t \cdot \frac{s}{\sqrt{n}} \\ &= 65 - 1.8331 \cdot \frac{0.7792}{\sqrt{10}} \\ &\approx 64.5483 \end{aligned} \qquad \text{(2pt)}$$

The critical region is therefore be given by $(-\infty; 64.5483]$. The sample mean (test (1pt) statistic) $\overline{x} = 63.75$ lies in the critical region, hence we reject the null hypothesis H_0. Based on the selected sample, there is sufficient evidence to believe that the new stealth design indeed reduces detectability. (1pt)

3d [4pt] Calculate the probability of a Type-II error β if the true reflected signal strength is actually normally distributed with $\mu = 64.5$ en $\sigma = 0.8$ dB (for a single observation).

Solution

We computed in the previous subquestion the critical region, which was equal to $(-\infty; 64.5483]$. Therefore, we need to compute the probability that given $\mu = 64$ en $\sigma = 0.8$, we get a value inside the acceptable region. (1pt)

In other words:

$$\begin{aligned}\beta &= P(\overline{X} < 64.5483 \mid \mu = 64.5)\\ &= \mathrm{normalcdf}(\text{lower} = -10^{99}; \text{upper} = 64.5483; \mu = 64.5; \sigma = \frac{0.8}{\sqrt{10}})\\ &\approx 0.5757\end{aligned}$$ (2pt)

So, $\beta \approx 0.5757$: a $57,57\%$ chance of accepting the null hypothesis while it is incorrect in reality. (1pt)

Problem 4 (20 points) Over a six-month period, NATO cyber experts monitor cyberattacks targeting Estonia. The following five types of attacks are recorded: phishing, malware injection, DDoS (Distributed Denial-of-Service), brute-force login attempts and supply chain exploits. The Estonian team observed 500 attacks in total. Furthermore, the expected distribution of cyberattack types is assessed based on global intelligence reports:

Attack type	Observed frequencies	Expected proportion
Phishing	90	20%
Malware injection	160	30%
DDoS	120	25%
Brute-force login attempts	70	15%
Supply chain exploits	60	10%

The cyber team would like to test whether the observed distribution of the types of incoming cyber attacks differs significantly from the expected distribution

4a [3pt] Which kind of hypothesis test do we need to perform. State the null hypothesis H_0 and the alternative hypothesis H_1 of this test.

Solution

Since we aim at testing whether the frequencies of a nominal variable follows a specific distribution, we need to perform a chi-square goodness-of-fit test. The null hypothesis H_0 and the alternative hypothesis H_1 of this test can be formulated as follows: (1pt)

H_0 : The distribution of attack types in Estonia follows
the expected proportions (20%, 30%, 25%, 15%, 10%). (1pt)

H_1 : The distribution of attack types in Estonia differs from
the expected proportions (20%, 30%, 25%, 15%, 10%). (1pt)

4b [3pt] Calculate the expected frequencies under the null hypothesis H_0.

Solution

We can evaluate the expected frequencies by taking the total of 500 observations and calculating frequencies based on the percentages. (1pt)

Attack type	Observed frequencies	Expected frequencies
Phishing	90	$20\% \cdot 500 = 100$
Malware injection	160	$30\% \cdot 500 = 150$
DDoS	120	$25\% \cdot 500 = 125$
Brute-force	70	$15\% \cdot 500 = 75$
Supply chain exploits	60	$10\% \cdot 500 = 50$
Total	500	500

(2pt)

4c [9pt] Perform the hypothesis test at a significance level $\alpha = 0.05$, and compute the p-value.

Solution

The test statistic for a chi-squared goodness-of-fit test are given by

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$ (1pt)

where O_i and E_i are respectively the observed and expected frequencies for category $i = 1, 2, 3, 4, 5$.

We can use the observed and expected frequencies to calculate the specific test statistic:

$$\begin{aligned}\chi^2 &= \frac{(90-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(120-125)^2}{125} + \frac{(70-75)^2}{75} + \frac{(60-50)^2}{50} \\ &= \frac{100}{100} + \frac{100}{150} + \frac{25}{125} + \frac{25}{75} + \frac{100}{50} \\ &= 1 + 0.6667 + 0.2 + 0.3333 + 2 \\ &= 4.2\end{aligned}$$ (4pt)

Since we have five categories, the number of degrees of freedom is one less, so df$=4$. Using the graphical calculator, we can compute the p-value of the test-statistic χ^2 as follows: (1pt)

$$p = P(X^2 \geq \chi^2 = 4.2) = \chi^2\text{cdf}(\text{lower} = 4.2; \text{upper} = 10^{99}; \text{df} = 4) \approx 0.3796$$ (3pt)

4d [5pt] State a conclusion for the hypothesis test in the original context of the problem and interpret it using data from the table.

Solution

The p-value of the test statistic χ^2 is larger than the significance level α, since $0.3796 > 0.05$. This means that based on the given sample we decide to accept (1pt) the null hypothesis H_0. There is insufficient reason to reject the claim that the (1pt) distribution of cyber attacks follows the given expected distribution. (1pt)

If we consider the table of observed and expected frequencies in subquestion (b), this makes sense as the observed and expected frequencies are relatively close to each other, and no surprisingly low or high values are observed. (2pt)

Problem 5 (18 points) In a test for the endurance of soldiers, research is conducted on the relationship between the load weight carried by the soldier (X) and the time needed for completing a 3 km speed march (Y). The following data were collected on 12 soldiers:

X	12	14	16	18	20	22	24	26	28	30	15	25
Y	16.75	16.29	17.97	19.78	17.65	18.15	21.37	20.65	19.3	21.31	16.05	18.55

The research team believes that there is a linear relationship between the load weight carried by the soldier and the time needed for completing the speed march.

5a [8pt] Calculate the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the slope and intercept of the linear regression line $Y = \beta_0 + \beta_1 X$, where Y is the speed march time, and X is the load weight carried.

Solution

We start by generating a table based on the data, the sample sums and means can then be used to calculate the least-squares estimated $\hat{\beta}_0$ and $\hat{\beta}_1$:

X	Y	XY	X^2	Y^2
12	16.75	201.0	144	280.5625
14	16.29	228.06	196	265.3641
16	17.97	287.52	256	322.9209
18	19.78	356.04	324	391.2484
20	17.65	353	400	311.5225
22	18.15	399.3	484	329.4225
24	21.37	512.88	576	456.6769
26	20.65	536.9	676	426.4225
28	19.30	540.4	784	372.49
30	21.31	639.3	900	454.1161
15	16.05	240.75	225	257.6025
25	18.55	463.75	625	344.1025
$\overline{X} = 20.8333$	$\overline{Y} = 18.6517$	$\overline{XY} = 396.5750$	$\overline{X^2} = 465.8333$	$\overline{Y^2} = 351.0376$

(4pt)

Using the standard formulas for the least-squares estimates, we get

$$\begin{aligned}\hat{\beta}_1 &= \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2} \\ &= \frac{396.575 - 20.833 \cdot 18.652}{465.833 - (20.833)^2} \\ &= \frac{7.999}{31.806} = 0.2515 \\ \hat{\beta}_0 &= \overline{Y} - \beta_1 \cdot \overline{X} \\ &= 18.652 - 0.2515 \cdot 20.833 \\ &= 13.4124. \end{aligned}$$ (3pt)

The equation of the regression line is thus equal to $Y = 13.4124 + 0.2515X$. (1pt)

5b [4pt] Interpret the slope β_1 of the regression model in the context of this exercise. What does the slope suggest about the relationship between load weight and speed march time?

Solution

In the context of this question, the slope β_1 of the regression tells us by how much more time it takes to finish a 3km speed march if the load increases by one kilogram. Since the least-squares estimate $\hat{\beta}_1$ of the slope is positive, this (1pt) suggests that the speed march time increases whenever the load weight increases. Specifically, for every additional kilogram of load weight, the predicted time to (2pt) complete the speed march increases by around a quarter of a minute (15 seconds). (1pt)

5c [6pt] Calculate the correlation coefficient $R(X, Y)$ between the load weight and speed march time. Based on the value of the correlation coefficient, what can you conclude about the strength and direction of the relationship between the two variables?

Solution

The correlation coefficient $R(X,Y)$ can be computed as follows:

$$\begin{aligned} R(X,Y) &= \frac{\overline{X \cdot Y} - \overline{X} \cdot \overline{Y}}{\sqrt{(\overline{X}^2 - \overline{X^2}) \cdot (\overline{Y}^2 - \overline{Y^2})}} \\ &= \frac{396.575 - 20.833 \cdot 18.652}{\sqrt{(20.833^2 - 465.833) \cdot (18.652^2 - 351.038)}} \\ &= \frac{7.999}{10.014} \\ &\approx 0.7987. \end{aligned}$$

(4pt)

The value of the correlation coefficient is positive and relative close to one, which suggests a strong positive correlation between load weight and speed march time. This makes sense, as speed marches are going to be harder once you have to carry a higher load weight with you. (2pt)