KASDI MERBAH UNIVERSITY, OUARGLA
DEPARTMENT OF COMPUTER SCIENCE AND IT
STATISTICS FOR DATA SCIENCE COURSE
M1 ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

MIDTERM EXAM CORRECTION

Exercise 1. **(6 pts)** A manufacturing company produced 3 batches, B_1, B_2, and B_3 of light bulbs during this week. The quality control team randomly selects samples of 10, 5, and 12 bulbs from the three batches, respectively, to test for defects. The historical defect rate for these bulbs is known to be 5%.

1. Calculate the probability that there are at least 2 defective light bulbs in a sample of 5 bulbs.
2. What is the average number of defective bulbs in the third sample?
   Let X be the total number of defective light bulbs in the three samples.
3. Determine the distribution and the parameters of X?
4. Use an appropriate approximation to find the probability that the total number of defective light bulbs in the three samples is at least equal to four.

Solution
Let X_i be the random variable denoting the number of defective light bulbs in the sample S_i **(0.5pt)**. We have

$$X_i \sim \mathcal{B}(n_i, p)$$

where $p = 0.05$ and $n_1 = 5, n_2 = 10, n_3 = 12$ and the distribution probability of X_i is

$$P(X_i = x) = \begin{cases} \binom{n}{x}(0.05)^x(0.95)^{n_i - x} & \text{if } x \in \{0, 1, 2, \dots, n_i\} \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(0.5pt)}$$

1. We have $P(X_2 \geq 2) = 1 - P(X_2 \leq 1)$ **(0.5pt)** where

   $$P(X_1 \leq 1) = P(X_1 = 0) + P(X_1 = 1) = 0.97741$$

   Therefore $P(X_2 \geq 2) = 0.0226$ **(0.5pt)**
2. The average number of defective light bulbs is $n_3 \times p = 12 \times 0.05 = 0.6$ **(0.5pt)**
3. Let X be the total number of defective light bulbs in the three samples. Thus $X = X_1 + X_2 + X_3$ **(0.75pt)**. Since $X_i \sim \mathcal{B}(n_i, 0.05)$ are independent, then $X \sim \mathcal{B}(n, 0.05)$ with $n = n_1 + n_2 + n_3 = 27$ **(0.5pt)**.
4. We have $n \geq 20$ and $p = 0.05 \leq 0.05$ **(0.5pt)**. We can thus approximate the binomial distribution of X by a Poisson distribution with the rate $\lambda = n \times p = 1.35$ **(0.75pt)**. The distribution of X is then as follows

   $$P(X = x) = \begin{cases} \mathrm{e}^{-1.35} \dfrac{(1.35)^x}{x!} & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(0.5pt)}$$

   Thus $P(X \geq 4) = 1 - P(X \leq 3) \approx 0.05$ **(0.5pt)**

Exercise 2. **(7 pts)** A machine, A, for checking computer chips uses an average of 65 milliseconds per check with a standard deviation of 4 milliseconds. It is assumed that check times are normally distributed and independent.

1. Calculate the probability that the check time does not exceed 60 milliseconds on machine A.
   A newer and cheaper machine B, potentially to be bought, uses on average 54 milliseconds per check.
2. If we know that in $84,13\%$ of cases, the check time for machine B, does not exceed 57 milliseconds, find its standard deviation.
3. What is the probability that the **time savings** per check using machine B instead of machine A is less than 10 milliseconds?
   The purchase of machine B depends on the check time of 100 chips, which is required to take less than 5.46 seconds in at least 95% of cases.
4. What decision can we make about the purchase of machine B?

Solution

1. Let X_A be the check time per one chip of machine A . $X_A \sim \mathcal{N}(65, 4)$. **(0.5pt)** Thus

$$P(X_A \leq 60) = P(\frac{X_A - 65}{4} \leq (\frac{60 - 65}{4}) = \Pi(-1.25) = 1 - \Pi(1.25) = 0.1056 \quad \textbf{(0.5pt)}$$

2. Let X_B be the check time per one chip of machine B, $X_B \sim \mathcal{N}(54, \sigma_B)$ **(0.25pt)**. It is given that $P(X_B \leq 57) = 0.8413$, thus

$$P(X_B \leq 57) = P(\frac{X_B - 54}{\sigma_B} \leq (\frac{57 - 54}{\sigma_B}) = \Pi(\frac{3}{\sigma_B}) = 0.8413$$

   From the Table, we can deduce that $\frac{3}{\sigma_B} = 1$ and thus $\sigma_B = 3$ **(0.75pt)**. Therefore, $X_B \sim \mathcal{N}(54, 3)$

3. We define T the random variable denoting the saving in check time between the two machines. Thus $T = X_A - X_B$ **(0.75pt)**. Since X_A and X_B are independent, we have $T \sim \mathcal{N}(\mu, \sigma)$ **(0.25pt)** with

$$\mu = E(T) = E(X_A - X_B) = E(X_A) - E(X_B) = 65 - 54 = 11 \quad \textbf{(0.75pt)}$$

   and

$$\sigma^2 = Var(T) = Var(X_A - X - B) = Var(X_A) + Var(X_B) = 16 + 9 = 25$$

   Thus $\sigma = 5$ **(0.75pt)** and $T \sim \mathcal{N}(11, 5)$.
   We are seeking $P(T \leq 10) = P(\frac{T - 11}{5} \leq -\frac{1}{5}) = 1 - \Pi(0.2) = 0.4207$ **(0.5pt)**

4. Let $S = \sum_{i=1}^{100} X_B^i$ be the check time of 100 chips for machine B **(0.5pt)**. $S \sim \mathcal{N}(100 \times 54, \sqrt{100} \times 3)$ **(0.5pt)**. To take a decision about the purchase of the machine B, we need to calculate $P(S \leq 5460)$ which is required to be at least 0.95. We have

$$P(S \leq 5460) = P(\frac{S - 5400}{30} \leq \frac{5460 - 5400}{30}) = \Pi(2) = 0.9792 > 0.95 \quad \textbf{(0.5pt)}$$

   Therefore, we can decide to buy the machine B **(0.5pt)**.

Exercise 3. **(7 pts)** An e-commerce company is interested in assessing customer satisfaction for a website layout. They collected feedback scores from a small sample of users who experienced the website layout. The satisfaction scores are as follows:

$$82, 85, 80, 78, 84, 88, 87, 90, 84, 89$$

We assume that the satisfaction score is a normal distribution with parameters μ and σ

1. Find the estimates for μ and σ
2. Determine a confidence interval for both parameters at a confidence level of 90%.
3. The company believes that the average satisfaction score of the customers about the website layout is 85. Perform a hypothesis test at a 5% significance level to determine if there is enough evidence to support this claim.

Solution

1. We have $A = \sum x_i = 847$ and $B = \sum x_i^2 = 71879$.
   - $\hat{\mu} = \overline{x} = \dfrac{A}{10} = 84.7$ **(0.5pt)**
   - $\hat{\sigma} = \sqrt{\dfrac{n}{n-1}}\sigma_{ech} = 3.9172$ **(0.5pt)**, with $\sigma_{ech} = \sqrt{\dfrac{B}{10} - A^2} = 3.71$ **(0.5pt)**

2. Since $n < 30$ and σ_{pop} is unknown, the random variable $T = \dfrac{\overline{X} - \mu}{\Sigma_{ech}/\sqrt{n-1}}$ is a student variable with $n-1$ degree of freedom. A confidence interval for μ is therefore of the form

$$\left[\overline{X} - t_{\alpha/2}\frac{\Sigma_{ech}}{\sqrt{n-1}}, \overline{X} + t_{\alpha/2}\frac{\Sigma_{ech}}{\sqrt{n-1}}\right] \quad \textbf{(0.5pt)}$$

   such that $P(|T| \geq t_{\alpha/2}) = 0.05$ From the table, we have $t_{\alpha/2} = 1.833$ **(0.5pt)**, therefore the 90% confidence interval for μ is $[82.43, 86.96]$ **(0.5pt)**.

   We find the 90% confidence interval for the variance, which is given in this case by $\left[\dfrac{n\Sigma_{ech}^2}{\chi_U^2}, \dfrac{n\Sigma_{ech}^2}{\chi_L^2}\right]$ **(0.5pt)**, since the random variable $Y = \dfrac{n\Sigma_{ech}^2}{\sigma_{pop}^2} \sim \chi_{n-1}^2$.
   χ_L^2 and χ_U^2 are such that $P(Y < \chi_L^2) = P(Y > \chi_U^2) = 0.05$. From the table we can deduce that $\chi_L^2 = 3,33$ and $\chi_U^2 = 16,92$ **(0.5pt)**. The confidence interval for the variance is then given by $[8.162, 41.47]$. The confidence interval for σ_{pop} can be deduced $[2.857, 6.44]$ **(0.5pt)**.

3. We have $\alpha = 0.05$. We construct the test as follows:
   - Formulation of hypotheses **(0.5pt)**: We have $\mathcal{H}_0 : \mu = \mu_0 = 85$ vs. $\mathcal{H}_1 : \mu \neq 85$.
   - Determination of the discriminant function **(0.5pt)**:
     Since the sample is small and the variance unknown, we define $T = \dfrac{\bar{X} - \mu_0}{\dfrac{\Sigma_{ech}}{\sqrt{n-1}}}$. $T \sim \mathcal{T}(n-1)$.
   - Finding the critical values $t_{\alpha/2}$ of T such that $P(|T| > t_{\alpha/2}) = 0.05$. From the table, we have $t_{\alpha/2} = 2.262$ **(0.5pt)**. therefore, the acceptance region is delimited by $[-2.262, 2.262]$.
   - Calculation of the value of T taken from the sample and conclusion of the test **(0.5pt)**: We have $T = \dfrac{84.7 - 85}{3.71/3} = -0.242 \in [-2.262, 2.262]$.
     Therefore, we accept the null hypothesis **(0.5pt)**.