# Probability and Statistics

Y-DATA School of Data Science

**P&P 6**

Due: 14.12.2022

- This final assignment contains one question from each topic.
- The work is individual, NOT in pairs.
- The weight of the assignment is 1/6 (i.e. no extra weight).

PROBLEM 1. During military training, three aircraft independently attacked a target, each exactly once. The first aircraft usually hits the target with a 0.6 probability, the second with a 0.4 probability, and the third with a 0.3 probability. After the training, it turned out that only one aircraft had hit the target. What is the probability that the 3rd one hit it?

Denote by $A_i$ the event that aircraft $i$ hits the target. It is given that $P(A_1) = 0.6$, $P(A_2) = 0.4$ and $P(A_3) = 0.3$. Let $X$ denote the number of hits on target. Therefore,

$$\begin{aligned} P(\text{aircraft 3 hit the target}|X=1) &= \frac{P(\text{aircraft 3 hit the target}, X=1)}{P(X=1)} \\ &= \frac{P(\text{aircraft 3 hit the target}, X=1)}{\sum_{i=1}^{3} P(\text{aircraft i hit the target}, X=1)} \\ &= \frac{0.4 \cdot 0.6 \cdot 0.3}{0.6 \cdot 0.6 \cdot 0.7 + 0.4 \cdot 0.4 \cdot 0.7 + 0.4 \cdot 0.6 \cdot 0.3} = 0.165 \end{aligned}$$

PROBLEM 2. A fair die is rolled until the total sum of points becomes 3 or more. Let $X$ denote the number of rolls until the above happens. Find the PMF and the CDF of $X$. Calculate $E(X)$ and $Var(X)$.

To compute the probability mass function, note that the maximum number of rolls is 3 (the minimal number in each roll is 1). The PMF is therefore given as follows:

$$P(X=1) = P(\text{first roll 3 or above}) = \frac{4}{6} = \frac{2}{3}$$

in the second case, we either get 1 in the first roll and 2-6 in the second, or we get 2 in the first and 1-6 in the second:

$$P(X=2) = \frac{1}{6} \cdot \frac{5}{6} + \frac{1}{6} \cdot 1 = \frac{11}{36}$$

in the third case, we need two get 1 in the first two rolls and anything in the third:

$$P(X=3) = \frac{1}{6} \cdot \frac{1}{6} \cdot 1 = \frac{1}{36}$$

Overall, the PMF is

$$P(X=k) = \begin{cases} 2/3, k=1 \\ 11/36, k=2 \\ 1/36, k=3 \end{cases}$$

and the CDF is

$$F_X(k) = \begin{cases} 0,\, k < 1 \\ 2/3,\, 1 \leq k < 2 \\ 35/36,\, 2 \leq k < 3 \\ 1,\, k \geq 3 \end{cases}$$

It is left to compute the expectation and the variance. To this end,

$$E(X) = \sum_{k=1}^{3} k \cdot P(X = k) = 1 \cdot 2/3 + 2 \cdot 11/36 + 3 \cdot 1/36 = 49/36 = 1.36$$

Recall that $Var(X) = E(X^2) - E^2(X)$ so,

$$E(X^2) = \sum_{k=1}^{3} k^2 \cdot P(X = k) = 1 \cdot 2/3 + 4 \cdot 11/36 + 9 \cdot 1/36 = 77/36$$
$$\Rightarrow Var(X) = 77/36 - (49/36)^2 = 0.286$$

PROBLEM 3. Let $X$ be a continuous random variable with PDF

$$f_X(x) = x/4, \text{ if } 1 < x \leq 3 \text{ and } 0 \text{ otherwise}$$

(1) Verify that $f_X(x)$ is indeed a PDF.
(2) Find $E(X)$.
(3) Let $Y = X^2$. Find $E(Y)$ and $Var(Y)$.
(4) Let $A = \{X \geq 2\}$. Find $P(A)$, $f_{X|A}(x)$ and $E(X|A)$.

(1) $f_X$ is a non-negative function, so it is left to verify that the integral over the support of $X$ is 1:

$$\int_1^3 f_X(x)dx = \int_1^3 x/4dx$$
$$= 1/4\left(x^2/2|_1^3\right) = (9/2 - 1/2)/4 = 1$$

and we are done.

(2)

$$E(X) = \int_1^3 x f_X(x)dx = 1/4 \int_1^3 x^2 dx$$
$$= 1/4\left(x^3/3|_1^3\right) = 13/6$$

(3) We know that

$$E(g(X)) = \int g(x) f_X(x)dx$$

So,

$$E(Y) = E(X^2) = \int_1^3 x^2 f_X(x)dx = 1/4 \int_1^3 x^3 dx = 1/4\left(x^4/4|_1^3\right) = 5$$

and

$$E(Y^2) = E(X^4) = \int_1^3 x^4 f_X(x)dx = 1/4 \int_1^3 x^5 dx = 1/4\left(x^6/6|_1^3\right) = 30.333$$

Therefore,

$$Var(Y) = 30.333 - 5^2 = 5.333$$

(4)

$$P(A) = P(X \geq 2) = \int_2^3 f_X(x)dx$$
$$= 1/4 \left(x^2/2|_2^3\right) = 5/8$$

We saw in class that

$$f_{X|A}(x) = \frac{f_X(x)1_A}{P(A)}$$

therefore,

$$f_{X|A}(x) = \frac{8}{5} \cdot \frac{1}{4} x 1_A = \frac{2}{5} x 1_{\{x \geq 2\}}$$

Lastly,

$$E(X|A) = \int x \cdot f_{X|A}(x)dx = \int_2^3 x \cdot \frac{2}{5} x dx$$
$$= 2/5 \left(x^3/3|_2^3\right) = 2.533$$

PROBLEM 4. Let $X_1, ..., X_n$ be an i.i.d. sample from the distribution with density

$$f_\theta(x) = \frac{x}{\theta} e^{-\frac{x^2}{2\theta}}, x \geq 0, \theta \geq 0$$

(1) Find the maximum likelihood estimator for $\theta$.
(2) Evaluate the MLE for the sample $(0.5, 0.5, 1)$

(1) We start by finding the likelihood function.

$$L(\theta; X) = \prod_{i=1}^n f_\theta(X_i) = \prod_{i=1}^n \frac{X_i}{\theta} e^{-\frac{X_i^2}{2\theta}} = \frac{\prod_{i=1}^n X_i}{\theta^n} e^{-\frac{1}{2\theta}\sum_{i=1}^n X_i^2}$$

the log-likelihood is then

$$\ell(\theta; X) = \log\left(\prod_{i=1}^n X_i\right) - n\log(\theta) - \frac{1}{2\theta}\sum_{i=1}^n X_i^2$$

To find the maximizer, we will take the derivative w.r.t. $\theta$ and equate it to 0.

$$\frac{d\ell(\theta; X)}{d\theta} = -\frac{n}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^n X_i^2 = 0$$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{1}{2n}\sum_{i=1}^n X_i^2$$

(2) For the sample $(0.5, 0.5, 1)$, we have $n = 3$ and

$$\hat{\theta}_{MLE} = \frac{1}{6}(2 \cdot 0.5^2 + 1) = 0.25$$

PROBLEM 5. It is known that an existing drug improves the health situation of 40% of the patients. Let $p$ be the probability of a positive effect of a **new** drug. After making sure that the side effects of the new drug are not different than those of the existing drug, the ministry of health wants to test whether the new drug has a bigger rate of success. To this end, a random sample of size $n$ is drawn and a hypothesis test is planned.

(1) Formulate the hypotheses $H_0$ and $H_1$.
(2) We reject the null if $\{\hat{p} > c\}$. Find $c$ using the normal approximation for significance level $\alpha = 0.05$.
(3) Compute the power of the test at the point $p = 0.55$ (you should get some function of $n$).
(4) Following the previous part, find the minimal sample size $n$ for which the power at the point $p = 0.55$ is at least 0.8.

(1) We have a one-sided hypothesis: $H_0 : p = 0.4$ vs. $H_1 : p > 0.4$.
(2) The estimator for the proportion is given by the sample mean $\bar{X}_n$ where each observation $X_i$ is $Ber(p)$. To find $c$ we solve the equation,

$$\alpha = P_{H_0}(\bar{X}_n > c) \approx 1 - \Phi\left(\sqrt{n}\frac{c-0.4}{\sqrt{0.4\cdot 0.6}}\right)$$
$$\Rightarrow c = 0.4 + z_{1-\alpha}\frac{\sqrt{0.4\cdot 0.6}}{\sqrt{n}} = 0.4 + 1.645\frac{0.49}{\sqrt{n}}$$

(3) The power of the test is the probability to correctly reject the null.

$$\pi = P_{H_1}\left(\bar{X}_n > 0.4 + 1.645\frac{0.49}{\sqrt{n}}\right) \approx 1 - \Phi\left(\sqrt{n}\frac{0.4 + 1.645\frac{0.49}{\sqrt{n}} - 0.55}{\sqrt{0.55\cdot 0.45}}\right)$$
$$= 1 - \Phi\left(1.62 - 0.3\sqrt{n}\right)$$

(4) We need to find $n$ for which

$$\pi \geq 0.8$$

That is,

$$1 - \Phi\left(1.62 - 0.3\sqrt{n}\right) \geq 0.8$$

Solving the inequality we get,

$$n \geq \left(\frac{1.62 - z_{1-0.8}}{0.3}\right)^2 = 67.328$$

That is, we should take $n > 67$.