Exercise Set 6 - Solution

1 The Binomial distribution and the grocer

a) X follows the binomial distribution with $N = 5$ and $p = 0.75$: $P(X = n) = \binom{N}{n} p^n (1-p)^{N-n}$

$X = n$	0	1	2	3	4	5
Probability $P(X = n)$	0.098%	1.465%	8.789%	26.367%	39.551%	23.730%

b) A client complains if $X \leq 3$, so the probability it happen is

$$p_{complain} = P(X \leq 3) = 1 - P(X = 4) - P(X = 5) = 36.719\%$$

c) If 100 clients buy the package, on average 36.719 will complain.

d) For a complaining customer, the grocer has received CHF 2.- , but has a cost of CHF 2.-, so there is no gain or loss. For non-complaining customers there is a profit of CHF 1.- however, so there is a profit overall.

To get the total gain per client, we can write :

$$\begin{aligned} \mu_{\text{gain}} &= p_{\text{complain}} \cdot \text{Gain(Complain)} + p_{\text{no complain}} \cdot \text{Gain(No Complain)} \\ &= p_{\text{complain}} \cdot (2 - 1 - 1) + p_{\text{no complain}} \cdot (2 - 1) \\ &= 0.63281 \approx 0.63 > 0 \end{aligned}$$

So the grocer makes a profit on average.

e) Now, for a complaint, there is a loss of CHF 2.- (two packages bought from the farmer, but no income).

The expected gain now becomes :

$$\begin{aligned} \mu_{\text{gain}} &= p_{\text{complain}} \cdot \text{Gain(Complain)} + p_{\text{no complain}} \cdot \text{Gain(No Complain)} \\ &= p_{\text{complain}} \cdot (2 - 1 - 1 - 2) + p_{\text{no complain}} \cdot (2 - 1) \\ &= -0.10157 \approx -0.10 < 0 \end{aligned}$$

So the grocer loses money *on average* although there is of course a certain probability that the grocer does still earn money for a finite number of clients.

2 Photodetector Efficiency

a) In this case, we are given the mean $\bar{X} = 1.3A/W$ of our sample (so only an estimate of the true mean μ of our process) and the true standard deviation $\sigma = 0.1\mu_0 = 0.13A/W$ of our process, so we use the **z-test**. Furthermore, we want to make sure our efficiency is higher than the reference mean $\mu_0 = 1.2A/W$, so we use a **one-sided z-test**.

b) It is always good practice to formulate our hypothesis. Here the hypothesis is that our efficiency is higher, so $H_1 : \mu > 1.2A/W$ and hence $H_0 : \mu \leq 1.2A/W$. We also choose a level of significance $\alpha = 0.01$.

We can calculate the z-value for our statistic : $Z(X) = \frac{\mu_0 - \bar{X}}{\sigma}\sqrt{4} = 1.54$. From here there are two ways of solving the problem : either compare the z-values, or compare the probabilities.

- Comparing the z-values. We calculate the z-value corresponding to a confidence level $\alpha = 0.01$, $z_{0.99} = 2.33$. This creates an exclusion interval $K = [2.33, +\infty[$.
  Here $Z(X) = 1.54 \notin K$ so we cannot refute H_0.
- Comparing the probabilities. From the z-value of our statistic, we can deduce a confidence level. Using the z-table, we have $z_{0.9382} = 1.54$, meaning we have a confidence level $\alpha_0 \approx 0.94 < 0.99$ for our true mean to actually be greater than μ_0.

In both cases, we conclude that our photodetector do not meet the standard with sufficient confidence level.

c) Here we want to know with which confidence level our value of the mean would be on the edge of refuting H_0. If we already computed the probability in exercise **b)**, we can already answer that $\alpha_{\text{limit}} \approx 0.94$. In this case, the probability that our probability is actually worse than the reference value is $P = 1 - 0.94 = 6\%$.

d) Now we want to know how many times we should do the experiment to be able to reject H_0 (assuming the sample mean stays the same). This means setting the condition :

$$\frac{\bar{X} - \mu_0}{\sigma}\sqrt{N} \geq z_{0.99} \Rightarrow N \geq (\frac{\sigma}{\bar{X} - \mu_0} z_{0.99})^2$$

Doing so we get $N \geq 9.25$, so we need to do at least 10 experiments.

3 Two different estimators of a macromolecule length measurement

a) The cumulative distribution function (i.e. the probability to find a value smaller or equal than x) is:

$$P(X \leq x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} c \cdot dx = \begin{cases} 0 & x \leq 0 \\ c \cdot x & 0 < x \leq \theta \\ c \cdot \theta & x > \theta \end{cases}$$

Since $P(X \leq \infty) = P(X \leq \theta)$ should be one for the probability to be normalized, we have to have $c = 1/\theta$.

b) The expectation value of the first estimator can be found by operations on the expectation operator. To find the expectation value of any measurement X_j, we take the integral over all values weighted by the cumulative distribution function, which just gives us $\theta/2$ as expected.

$$\mathbb{E}(\theta_1) = \mathbb{E}\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = 2 \cdot \mathbb{E}(X_j) = 2\int_0^\theta \frac{x}{\theta}dx = \theta$$

For the second one, we use the density function, which is the derivative of the cumulative distribution.

$$\begin{aligned} F_{\theta_2}(t) &= P(\max_i X_i \leq t) = P(X_1 \leq t, ..., X_n \leq t) = P(X_j \leq t)^n = \left(\frac{t}{\theta}\right)^n \\ f_{\theta_2}(t) &= \frac{d}{dt}F_{\theta_2}(t) = n\frac{t^{n-1}}{\theta^n} \\ \mathbb{E}(\theta_2) &= \mathbb{E}(\max_i X_i) = \int_0^\theta f_{\theta_2}(x)x \cdot dx = n\int_0^\theta \frac{x^n}{\theta^n}dx = \frac{n}{n+1}\theta \end{aligned}$$

So the second estimator is biased. A non-biased estimator would be $\theta_2^* = \frac{n+1}{n}\theta_2$.

c) The mean square errors are just the variances because the bias is zero:

$$\begin{aligned}
MSE(\theta_1) &= \mathbb{E}\left(\left(\frac{2}{n}\sum_{i=1}^{n} X_i - \theta\right)^2\right) = \frac{4}{n^2}\mathbb{E}\left(\left(\sum_{i=1}^{n} X_i\right)^2\right) - \frac{4\theta}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) + \theta^2 \\
&= \frac{4}{n^2}\left(n \cdot \mathbb{E}(X_j^2) + n(n-1)\cdot \mathbb{E}(X_j)^2\right) - \frac{4\theta}{n}\frac{n\theta}{2} + \theta^2 \\
&= \frac{4}{n}\left(\mathbb{E}(X_j^2) + (n-1)\frac{\theta^2}{4}\right) - \theta^2 = \frac{4}{n}\int_0^\theta \frac{x^2}{\theta} dx - \frac{\theta^2}{n} \\
&= \frac{\theta^2}{3n} \\
MSE(\theta_2^*) &= \mathbb{E}\left(\left(\frac{n+1}{n}\theta_2 - \theta\right)^2\right) = \frac{(n+1)^2}{n^2}\mathbb{E}\left(\theta_2^2\right) - \frac{2\theta(n+1)}{n}\mathbb{E}(\theta_2) + \theta^2 \\
&= \frac{(n+1)^2}{n^2}\int_0^\theta f_{\theta_2}(x)x^2 \cdot dx - \theta^2 = \frac{(n+1)^2}{n}\int_0^\theta \frac{x^{n+1}}{\theta^n} dx - \theta^2 \\
&= \theta^2\left(\frac{(n+1)^2}{n(n+2)} - 1\right) \\
&= \frac{\theta^2}{n(n+2)}
\end{aligned}$$

d) θ_2^* is the best estimator for θ, because its mean square error is smaller. (Technically they are the same for n=1, but clearly for 1 measurement the variance it not well defined). Importantly, the MSE decreases quadratically with the number of measurements, rather than just linearly.