Introduction to Data Science (1MS041)
Uppsala University – Autumn 2024
Report for Assignment 2

Per Vincent Ankarbåge, Jonathan Falk, Madhur Gupta,
Henrik Jonasson, Adam Rokah

November 8, 2024

All group members contributed equally by individually taking on assigned problems (each name is specified in the title for their respective question), following up by collective discussion for fine-tuning the solutions and reporting.

1 Madhur

The conditional density of $Y|X$ is given by:

$$f_{Y|X}(y,x) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad \lambda(x) = \exp(\alpha \cdot x + \beta),$$

where α is a vector (slope) and β is a number (intercept).

For a single observation (y_i, x_i), the log-likelihood function based on the Poisson distribution is:

$$\log f_{Y|X}(y_i, x_i) = y_i \log(\lambda(x_i)) - \lambda(x_i) - \log(y_i!).$$

For n observations $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, the total log-likelihood is the sum of individual log-likelihoods:

$$\sum_{i=1}^{n} \log f_{Y|X}(y_i, x_i) = \sum_{i=1}^{n} \left(y_i \log(\lambda(x_i)) - \lambda(x_i) - \log(y_i!) \right).$$

Next, take the negative of the log-likelihood. Thus, the loss function we want to minimize is:

$$-\sum_{i=1}^{n} \left(y_i \log(\lambda(x_i)) - \lambda(x_i) - \log(y_i!) \right).$$

Simplifying this expression, we get:

$$-\sum_{i=1}^{n} y_i \log(\lambda(x_i)) + \sum_{i=1}^{n} \lambda(x_i) + \sum_{i=1}^{n} \log(y_i!).$$

Since $\log(y_i!)$ is constant with respect to α and β, it does not affect the minimization. Therefore, we can ignore this term, and the loss function $Loss(\alpha, \beta)$ that we need to minimize becomes:

$$Loss(\alpha, \beta) = -\sum_{i=1}^{n} y_i \log(\lambda(x_i)) + \sum_{i=1}^{n} \lambda(x_i).$$

The factorial term $\log(y_i!)$ is unnecessary for the optimization since it does not affect the minimization with respect to the parameters.

2 Per

2.1 Finding the distribution function

First we find the distribution function of $\hat{\theta}$.

For random variables X_i drawn from Uniform$(0, \theta)$ the cummulative distribution function (CDF) is given by

$$F_{x_i}(x) = P(X_i < x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{x}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 1, & \text{if } x > \theta. \end{cases}$$

The CDF $F_{\hat{\theta}}$ of the maximum $\hat{\theta} = \max(X_1, ..., X_n)$ can be derived by the CDF of these random variables F_{x_i} and the fact that $X_1, \ldots, X_n$ are IID.

$$\begin{aligned} F_{\hat{\theta}}(x) &= P(\hat{\theta} \leq x) \\ &= P(\max(X_1, X_2, \ldots, X_n) \leq x) \\ &= P(X_1 \leq x \cap X_2 \leq x \cap \cdots \cap X_n \leq x) \\ &= P(X_1 \leq x)^n \quad \text{(since the } X_i \text{ are IID we simply multiply the probabilites} \\ &= \left(\frac{x}{\theta}\right)^n, \quad 0 \leq x \leq \theta. \end{aligned}$$

2.2 Finding the Bias

Recall that the bias for a point estimator is defined as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

.

Let's first examine $\mathbb{E}[\hat{\theta}]$. Since $\hat{\theta}$ is a continuous random variable (uniform), the expected value of $\hat{\theta}$ is the integral of x weighted by its probability $f_{\hat{\theta}}(x)$:

$$\mathbb{E}[\hat{\theta}] = \int_0^\theta x \cdot f_{\hat{\theta}}(x)\, dx.$$

.

Now, to get $f_{\hat{\theta}}$ we differentiate $F_{\hat{\theta}}$ with respect to x.

$$f_{\hat{\theta}} = \frac{d}{dx}\left(\frac{x^n}{\theta^n}\right) = n\frac{x^{n-1}}{\theta^n}$$

Thus we have,

$$\mathbb{E}[\hat{\theta}] = \int_0^\theta n \cdot \frac{x^n}{\theta^n}\, dx = \frac{n}{n+1}\theta$$

.

Plugging it into the definition of bias we have,

$$\text{Bias}(\hat{\theta}) = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}$$

.

2.3 Finding the standard error

Recall the definition of the standard error for a point estimator

$$\text{se} = \sqrt{\text{Var}(\hat{\theta})}$$

.

Let's first evaluate the variance. Using the previous method to get $\mathbb{E}[\hat{\theta}^2]$ and our previous result for $\mathbb{E}[\hat{\theta}]$ we get that

$$\begin{aligned}
\text{Var}(\hat{\theta}) &= \mathbb{E}[\hat{\theta}^2] - \left(\mathbb{E}[\hat{\theta}]\right)^2 \\
&= \theta^2 \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) \\
&= \theta^2 \cdot n \left(\frac{1}{n+2} - \frac{n}{(n+1)^2}\right) \\
&= \theta^2 \cdot n \cdot \frac{(n+1)^2 - n(n+2)}{(n+2)(n+1)^2} \\
&= \theta^2 \cdot n \cdot \frac{1}{(n+2)(n+1)^2} \\
&= \frac{n\theta^2}{(n+1)^2(n+2)}.
\end{aligned}$$

Now taking the square root to get the standard error.

$$\begin{aligned}
\sqrt{\text{Var}(\hat{\theta})} &= \sqrt{\frac{n\theta^2}{(n+1)^2(n+2)}} \\
&= \theta \cdot \sqrt{\frac{n}{(n+1)^2(n+2)}} \\
&= \frac{\theta\sqrt{n}}{(n+1)\sqrt{n+2}}.
\end{aligned}$$

.

2.4 Finding the Mean Square Error

To find the MSE of the point estimator, we simply plug in our previous results into the definition of MSE and simplify the results.

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{Var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}))^2 \\
&= \frac{n\theta^2}{(n+1)^2(n+2)} + \left(-\frac{\theta}{n+1}\right)^2 \\
&= \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{\theta^2}{(n+1)^2} + \frac{\theta^2(n+2)}{(n+1)^2(n+2)} \\
&= \frac{n\theta^2 + \theta^2 n + 2\theta^2}{(n+1)^2(n+2)} \\
&= \frac{2\theta^2 n + 2\theta^2}{(n+1)^2(n+2)} \\
&= \frac{2\theta^2(n+1)}{(n+1)^2(n+2)} \\
&= \frac{2\theta^2}{(n+1)(n+2)}.
\end{aligned}$$

3 Jonathan

We are given a density function

$$p(x) = \frac{1}{2}\cos x, \ -\frac{\pi}{2} < x < \frac{\pi}{2},$$

of a continuous distribution. Part a) asks us to find the distribution function F, which is given by

$$F(x) = \int_{-\frac{\pi}{2}}^{x} p(x)dx = \frac{1}{2}\int_{-\frac{\pi}{2}}^{x} \cos x dx = \frac{1}{2}(\sin x + 1) = \frac{1+\sin x}{2}.$$

Part b) asks us to find the inverse distribution F^{-1}. This is given by

$$F^{-1}(F(x)) = x,$$

$$F^{-1}\left(\frac{1+\sin x}{2}\right) = x.$$

We can see that $F^{-1}(x) = \arcsin 2x - 1$, since

$$F^{-1}(F(x)) = \arcsin(2F(x)-1) = \arcsin\left(2\left(\frac{1+\sin x}{2}\right) - 1\right) = \arcsin\left(\sin(x)\right) = x.$$

For task c), in order to sample $p(x)$ using the Accept-Reject sampler we are tasked with finding a density $g(x)$ such that $p(x) \leq Mg(x)$ for some $M > 0$. We can simply let $g(x) = C$ for some $C > 0$ and $-\frac{\pi}{2} < x < \frac{\pi}{2}$. This guarantees that $g(x) \geq 0$ for all x. In order to find $g(x)$, we notice that we must have

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} g(x)dx = 1$$

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} C dx = 1$$

$$C\left(\frac{\pi}{2} - \frac{-\pi}{2}\right) = 1$$

$$C = \frac{1}{\pi}.$$

This means that $g(x) = \frac{1}{\pi}$. In order to find M, we see that $p(x) = \frac{1}{2}\cos x \leq \frac{1}{2}$ for $-\frac{\pi}{2} < x < \frac{\pi}{2}$. Thus, $\frac{1}{2} = M\frac{1}{\pi}$, which gives us $M = \frac{\pi}{2}$.

4 Henrik

Recall that a stochastic process $X_0, X_1, ..., X_n$ is a Markov chain if for any $t \in 0..n$ the following holds:

$$\mathbb{P}(X_t = x | X_0, X_1, .., X_{t-1}) = \mathbb{P}(X_t = x | X_{t-1})$$

That is, the probability that $X_t = x$ only depends on the previous state X_{t-1}. To verify that $X_t = \max\{Y_1, ..., Y_t\}$ is a Markov chain (where $X_0 = 0$ and $Y_1, .., Y_n$ are IID discrete random variables), we simply note that:

$$\max\{Y_1, ..., Y_t\} = \max\{\max\{Y_1, ...Y_{t-1}\}, Y_t\} = \max\{X_{t-1}, Y_t\}$$

Which yields that:

$$\mathbb{P}(X_t = x|X_0, X_1, .., X_{t-1}) = \mathbb{P}(\max\{Y_1, ..., Y_t\} = x) =$$

$$\mathbb{P}(\max\{X_{t-1}, Y_t\} = x) = \mathbb{P}(X_t = x|X_{t-1})$$

Furthermore, our Markov chain is homogeneous. To see this, note that:

$$\mathbb{P}(X_t = y|X_{t-1} = x) = \begin{cases} \mathbb{P}(Y_t = y), & \text{if } y > x, \\ \mathbb{P}(Y_t \leq y), & \text{if } y = x, \\ 0, & \text{if } y < x. \end{cases} \tag{1}$$

In all these cases, the probability for $\mathbb{P}(X_t = y|X_{t-1} = x)$ is independent of time for any t for which $X_{t-1} = x$, i.e. our process satisfies the following definition of a homogeneous Markov chain:

$$\mathbb{P}(X_t = y|X_{t-1} = x) = \mathbb{P}(X_s = y|X_{s-1} = x) \text{ for any } s, t \in 0..n$$

Thus we can define our Markov chain by a transition matrix P where $P_{xy} = \mathbb{P}(X_t = y|X_{t-1} = x)$. Using equation (1) together with $\mathbb{P}(Y_t = 0) = 0.1$, $\mathbb{P}(Y_t = 1) = 0.3$, $\mathbb{P}(Y_t = 2) = 0.2$ and $\mathbb{P}(Y_t = 3) = 0.4$ we now get:

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.2 & 0.4 \\ 0 & 0.4 & 0.2 & 0.4 \\ 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

5 Adam

To estimate the quantile x_p of an unknown distribution F, we start by understanding that x_p represents the p-quantile of F. This means that x_p is the value such that:

$$F(x_p) = p.$$

In other words, the probability that a random variable X drawn from F is less than or equal to x_p is exactly p. Since we do not know F, we use the empirical distribution function $\hat{F}_n$, which is based on our sample, as an approximation.

To approximate x_p, we use the empirical quantile $\hat{x}_p$, which is defined by:

$$\hat{F}_n(\hat{x}_p) \approx p.$$

This empirical quantile serves as an estimate of the true quantile x_p.

Now, we want to understand the error in this approximation. The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality provides a way to bound the maximum difference between the empirical distribution $\hat{F}_n$ and the true distribution F with high probability. Specifically, for any $\epsilon > 0$, the DKW inequality states:

$$P\left(\sup_x \left|\hat{F}_n(x) - F(x)\right| \leq \epsilon\right) \geq 1 - \alpha,$$

where $\epsilon = \sqrt{\frac{\ln(2/\alpha)}{2n}}$ for a chosen confidence level $1 - \alpha$.

This inequality tells us that, with probability $1 - \alpha$, the true CDF $F(x)$ is within ϵ of the empirical CDF $\hat{F}_n(x)$ for all x. Applying this to the quantile x_p, we can say:

$$\hat{F}_n(x_p) - \epsilon \leq p \leq \hat{F}_n(x_p) + \epsilon.$$

Thus, the value x_p, which satisfies $F(x_p) = p$, is likely to lie between the values of x where the empirical distribution $\hat{F}_n(x)$ equals $p - \epsilon$ and $p + \epsilon$.

To formalize this, let $x_{p,\text{lower}}$ be the value such that $\hat{F}_n(x_{p,\text{lower}}) = p - \epsilon$, and let $x_{p,\text{upper}}$ be the value such that $\hat{F}_n(x_{p,\text{upper}}) = p + \epsilon$. Then, with probability $1 - \alpha$, we have:

$$x_p \in [x_{p,\text{lower}}, x_{p,\text{upper}}].$$

This interval $[x_{p,\text{lower}}, x_{p,\text{upper}}]$ gives us a confidence interval for the true quantile x_p, based on the sample and the chosen confidence level $1 - \alpha$.