MAP 531: Homework

Paul-Antoine GIRARD & Adrien TOULOUSE

Problem 1: Estimating parameters of a Poisson distribution to model the number of goals scored in football

We recall that the Poisson distribution with parameter $\theta > 0$ has a pdf given by $(p(\theta, k), k \in \mathbb{N})$ w.r.t the counting measure on $\mathbb{N}$:

$$p(\theta, k) = e^{-\theta}\frac{\theta^k}{k!}$$

Question 1

The poisson distribution is a discrete distribution since it has a countable number of possible values ($\mathbb{N}$).

In statistics, we use this distribution to compute the probability of a given number of (rare) events in a time period or the probability of a discrete waiting time until the next event (eg. number of minutes).

For example a poisson distribution can model:

- The number of patients arriving in an emergency room between 9 and 10am.
- The number of minutes we wait a bus at the bus stop.
- In quality control, the number of manufacturing defect.

Question 2

We assume that $\mathbb{X}$ follows a Poisson distribution with parameter $\theta > 0$.

$$\mathbb{E}[\mathbb{X}] = \sum_{i=0}^{\infty}(i * p(\theta, i)) = \sum_{i=0}^{\infty}(i * e^{-\theta}\frac{\theta^i}{i!}) = \theta * e^{-\theta}\sum_{i=1}^{\infty}(\frac{\theta^{i-1}}{(i-1)!}) = \theta * e^{-\theta}\sum_{i=0}^{\infty}(\frac{\theta^i}{i!}) = \theta * e^{-\theta} * e^{\theta} = \theta$$

$$\mathbb{E}[\mathbb{X}^2] = \sum_{i=0}^{\infty}(i^2 * p(\theta, i)) = \sum_{i=0}^{\infty}(i^2 * e^{-\theta}\frac{\theta^i}{i!}) = \theta * e^{-\theta}\sum_{i=1}^{\infty}(i\frac{\theta^{i-1}}{(i-1)!}) = \theta * e^{-\theta}\sum_{i=0}^{\infty}((i+1)\frac{\theta^i}{i!})$$

$$= \theta * e^{-\theta}[\sum_{i=0}^{\infty}(i\frac{\theta^i}{i!}) + \sum_{i=0}^{\infty}(\frac{\theta^i}{i!})] = \theta * e^{-\theta}[\theta * e^{\theta} + e^{\theta}] = \theta(\theta + 1)$$

$$\mathbb{V}(\mathbb{X}) = \mathbb{E}[\mathbb{X}^2] - \mathbb{E}[\mathbb{X}]^2 = \theta(\theta + 1) - \theta^2 = \theta$$

Question 3

We are provided with n independent observations of a Poisson random variable of parameter $\theta \in \Theta = \mathbb{R}_+^*$.
Our observations are $X_k \sim Pois(\theta), \forall k \in 1, ..., n$.
The corresponding statistical model is

$$\mathbb{M} = \{p(. \mid \theta),\ \theta \in \Theta\}$$

We are trying to estimate the parameter θ.

Question 4

The likelihood function is the function on θ that makes our n observations most likely.

$$l(\theta) = \prod_{k=1}^{n} p(\theta, x_k) = \prod_{k=1}^{n} e^{-\theta} \frac{\theta^{x_k}}{x_k!}, with\ x_k \in \mathbb{N}, \forall k \in 1, ..., n$$

$$L(\theta) = log(l(\theta)) = \sum_{k=1}^{n} (-\theta + x_k log(\theta) - log(x_k!)) = -n\theta + log(\theta) \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} log(x_k!)$$

By derivating with respect to θ, we have:

$$L'(\theta) = -n + \frac{\sum_{k=1}^{n} x_k}{\theta}$$

Then, we set this derivative equal to zero to obtain a critical point:

$$L'(\theta) = 0 \Leftrightarrow -n + \frac{\sum_{k=1}^{n} x_k}{\theta} = 0 \Leftrightarrow \hat{\theta} = \overline{x}$$

and this critical point is a local maximum, and we will assume that it is also a global maximum of the likelihood function:

$$L''(\theta) = -\frac{\sum_{k=1}^{n} x_k}{\theta^2} < 0$$

So, the maximum likelihood estimator is:

$$\hat{\theta}_{MLE} = \overline{x}$$

Question 5

We have that:

$$\mathbb{E}[\overline{x}] = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[x_k] = \mathbb{E}[x_1] = \theta$$

$$\mathbb{V}(\overline{x}) = \frac{1}{n^2} \sum_{k=1}^{n} \mathbb{V}(x_k) = \frac{1}{n} \mathbb{V}[x_1] = \frac{\theta}{n}$$

Applying the central limit theorem, we have that $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$ converges towards a Gaussian $\mathcal{N}(0, \theta)$.

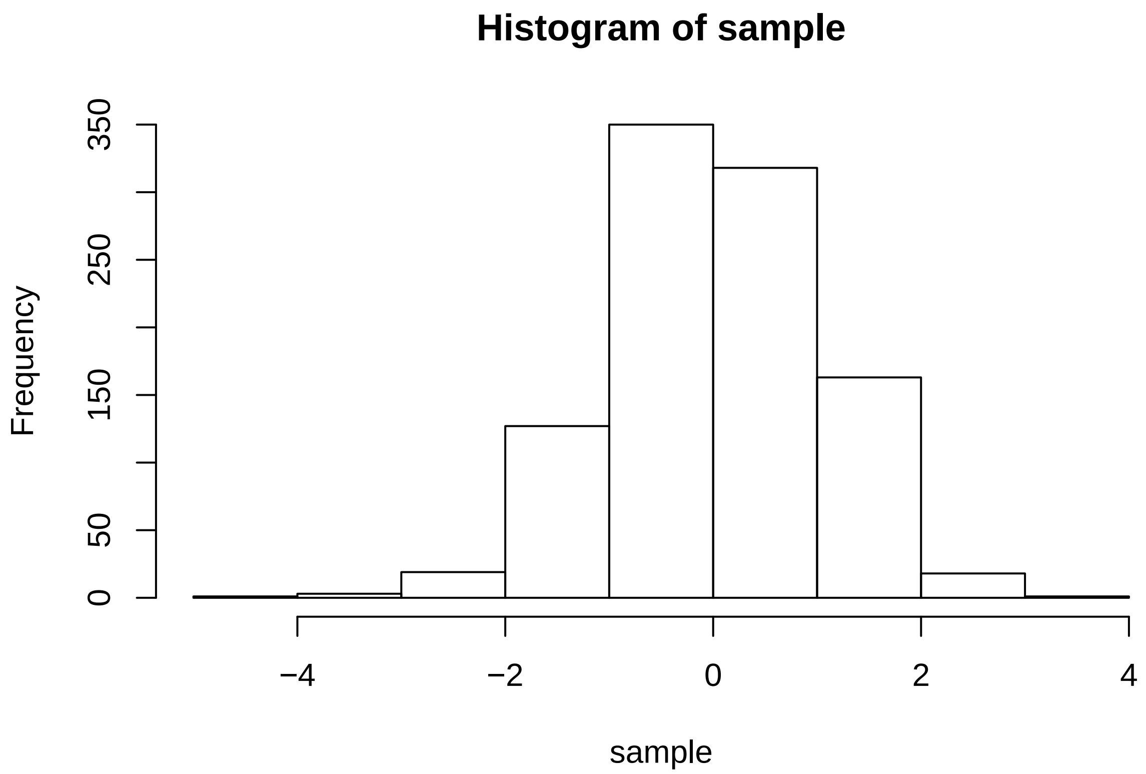
Question 6

By continuous mapping, $\sqrt{\hat{\theta}_{MLE}}$ converges in probability towards $\sqrt{\theta}$. Then, by Slutsky's theorem, we have that $\sqrt{n}\frac{(\hat{\theta}_{MLE} - \theta)}{\sqrt{\hat{\theta}_{MLE}}}$ converges in law towards a gaussian $\mathcal{N}(0, 1)$.

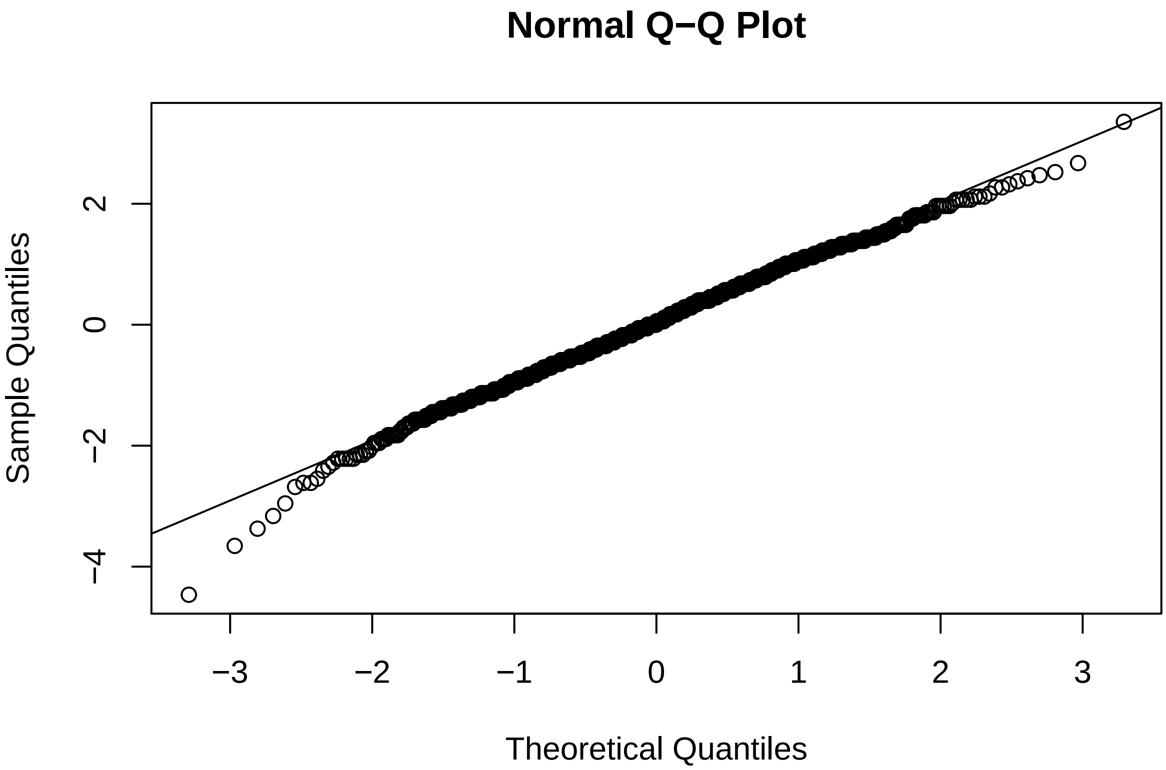
Let's check this result in R by simulating 1000 times our random variable $\sqrt{n}\frac{(\hat{\theta}_{MLE} - \theta)}{\sqrt{\hat{\theta}_{MLE}}}$ with a sample size of 100:

```r
Nattempts = 1000
nsample = 100
theta = 3
sample = rep(0, 1000)
for (i in 1:Nattempts)  # can be written without the for loop (nicer) !
{poisson_sample = rpois(nsample, theta)
  sample[i] = sqrt(nsample) * (mean(poisson_sample) - theta) / sqrt(mean(poisson_sample))
}

hist(sample)
```



```r
qqnorm(sample)
qqline(sample)
```



Question 7

Let Z_n be our random variable, so that $Z_n = \sqrt{n}\frac{(\hat{\theta}_{MLE}-\theta)}{\sqrt{\hat{\theta}_{MLE}}}$

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z_n \leq z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow \mathbb{P}(-z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_{MLE}}{n}} \leq \hat{\theta}_{MLE}-\theta \leq z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_{MLE}}{n}}) = 1-\alpha$$

For $\alpha \in (0,1)$, an asymptotic confidence interval for θ of level α is therefore :

$$[\hat{\theta}_{MLE} - z_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}};\ \hat{\theta}_{MLE} + z_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}}]$$

Question 8

We apply the δ-method with $g(x) = 2 \times \sqrt{x}$ We have: $g'(x) = \frac{1}{\sqrt{x}}$
So,

$$\sqrt{n}(\hat{\theta}_{MLE}-\theta) \xrightarrow{d} \mathcal{N}(0,\ g'(\theta)^2 \times \theta) \Leftrightarrow \sqrt{n}(\hat{\theta}_{MLE}-\theta) \xrightarrow{d} \mathcal{N}(0,1)$$

Question 9

Let Z_n be our random variable, so that $Z_n = \sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta})$

We know that $Z_n \overset{d}{\to} \mathcal{N}(0,1)$

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z_n \leq z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow \mathbb{P}(-\frac{z_{1-\alpha/2}}{2\sqrt{n}} \leq \sqrt{\hat{\theta}_{MLE}} - \sqrt{\theta} \leq \frac{z_{1-\alpha/2}}{2\sqrt{n}}) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}(\sqrt{\hat{\theta}_{MLE}} - \frac{z_{1-\alpha/2}}{2\sqrt{n}} \leq \sqrt{\theta} \leq \sqrt{\hat{\theta}_{MLE}} + \frac{z_{1-\alpha/2}}{2\sqrt{n}}) = 1-\alpha$$

For $\alpha \in (0,1)$, an asymptotic confidence interval for θ of level α is therefore :

$$[\hat{\theta}_{MLE} - z_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}};\ \hat{\theta}_{MLE} + z_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}}]$$

For $\alpha \in (0,1)$, an asymptotic confidence interval for θ of level α is therefore :

$$[**]$$

Question 10

Based on the first moment of a poisson distribution, we easily have that:

$$\hat{\theta}_{MME} = \overline{x}$$

We then remark that $\hat{\theta}_{MME} = \hat{\theta}_{MLE}$

Based on the second moment of a poisson distribution, we have:

$$n^{-1}\sum_{k=1}^{n} X_k^2 = \hat{\theta}_2(\hat{\theta}_2 + 1)$$

Let's define the function $h(x) = x(x+1)$
Its inverse on $\mathbb{R}_+^*$ is $h^{-1} = \frac{1}{2} - 1 + \sqrt{4x+1})$ and then we have that:

$$\hat{\theta}_2 = \frac{1}{2}[-1 + \sqrt{(4n^{-1}\sum_{k=1}^{n} X_k^2) + 1}]$$

Question 11

$\mathbb{E}(\hat{\theta}_{MLE}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)$ by linearity of the expectation $\mathbb{E}(\hat{\theta}_{MLE}) = \frac{1}{n} * n\theta = \theta$

Therefore, $\hat{\theta}_{MLE}$ is an unbiased estimator of θ, ie. $b_\theta{}^*(\hat{\theta}_{MLE}) = 0$

$\mathbb{V}(\hat{\theta}_{MLE}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i)$ by independance of the X_i $\mathbb{V}(\hat{\theta}_{MLE}) = \frac{1}{n^2} * n\theta = \frac{\theta}{n}$

The quadratic risk Q is given by : $Q = b_\theta{}^*(\hat{\theta}_{MLE}) + \mathbb{V}^*(\hat{\theta}_{MLE}) = 0 + \frac{\theta}{n} = \frac{\theta}{n}$

Question 12

$\hat{\theta}_{MLE}$ is an unbiased estimator so the Cramer-Rao bound is given by:

$$\frac{1}{I_n(\theta^*)} = \frac{1}{\mathbb{E}(-L''(\theta^*))}$$

$$L'(\theta^*) = -n + \frac{\sum_{i=1}^n x_k}{\theta}$$

$$-L''(\theta^*) = \frac{\sum_{i=1}^n x_k}{\theta^2}$$

Therefore,

$$\mathbb{E}(-L''(\theta^*)) = \frac{\sum_{i=1}^n \mathbb{E}(x_k)}{\theta^2} = \frac{n}{\theta}$$

Finally,

$$\frac{1}{I_n(\theta^*)} = \frac{\theta}{n} = \mathbb{V}(\hat{\theta}_{MLE})$$

We can conclude that our estimator $\hat{\theta}_{MLE}$ is efficient.

Question 13

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \theta + \theta - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1}^n [(X_i - \theta)^2 + (\theta - \overline{X_n})^2 + 2(X_i - \theta)(\theta - \overline{X_n})]$$

$$= \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 + (\theta - \overline{X_n})^2 + \frac{2}{n}(\theta - \overline{X_n})\sum_{i=1}^n (X_i - \theta) = \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 + (\theta - \overline{X_n})^2 + 2(\theta - \overline{X_n})(\overline{X_n} - \theta)$$

$$= \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - (\theta - \overline{X_n})^2$$

Question 14

$$\mathbb{E}(\theta - \overline{X_n})^2 = \mathbb{E}(\theta^2 - 2\theta\overline{X_n} + \overline{X_n}^2) = \theta^2 - 2\theta\mathbb{E}(\overline{X_n}) + \mathbb{E}(\overline{X_n})^2$$

$$= -\theta^2 + \mathbb{V}(\overline{X_n}) + \mathbb{E}(\overline{X_n})^2 = -\theta^2 + \frac{\theta}{n} + \theta^2 = \frac{\theta}{n}$$

$$\mathbb{E}(\hat{\theta}_2) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - (\theta - \overline{X_n})^2)$$

$$= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i - \theta)^2 - \mathbb{E}(\theta - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{V}(X_i) - \frac{\theta}{n} = \theta(1 - \frac{1}{n})$$

Therefore the bias is,

$$b_{\hat{\theta}_2} = \frac{\theta}{n}$$

We can get an unbiased estimator $\hat{\theta}_3$ by defining $\hat{\theta}_3 = \hat{\theta}_2 * (1 - \frac{1}{n})^{-1}$