# 2021F_Week03

September 10, 2021

## 1 Week03

The goals for week03 are to use the maximum likelihood approach to estimate parameters for a model and to characterize the variability in parameters using the Fisher information.

We will learn - The Fisher Information for one and N data points - The Fundemental Theorum of Maximum Likelihood - How to compute Confidence Intervals for parameters

### 1.0.1 Model setup

Suppose we have a data vector that contains $N$ datapoints

$$x = [x_1, x_2, \cdots, x_N] \tag{1}$$

We will assume each data point $x_i$ was generated from a corresponding random variable $X_i$, or

$$X_i \sim x_i \text{ for } i = 1 \text{ to } N \tag{2}$$

To model our data, we can further assume our random variables $X_i$ are **independent and identically distributed**

### 1.0.2 iid

We will assume every pair of random variables are *indepednent*, that is,

$$p(X_i|X_j = x) = p(X_i) \tag{3}$$

We also further assume that every random variable $X_i$ follows the same distribution with the same parameters. This is the identical distribution assumption and we often say " random variables $X_1, X_2, \cdots X_n$ are identically distributed" or that

$$X_i \sim f(x|\theta) \tag{4}$$

where $f$ is a probability density (or mass) function and $\theta$ is a set of corresponding parameters. Here, each random variable follows the *same* probability density/mass function with the *same* set of parameters.

With these assumptions we were able to simplify our the likelihood and so loglikelihood. Our likelihood $\mathcal{L}$ simplified to

$$\begin{aligned}
\mathcal{L}(\theta) &= p(x_1, x_2, x_3, \cdots, x_n|\theta) && (5)\\
&= p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)\cdots p(x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \qquad \text{Independence} && (6)\\
&= \prod_{i=1}^{n} f(x_i|\theta) \qquad \text{identically distributed} && (7)
\end{aligned}$$

and so our logliklihood was

$$\ell\ell(\theta) = \log[\mathcal{L}(\theta)] = \sum_{i=1}^{n} \log[f(x_i|\theta)] \tag{8}$$

We can maximize the above function for our parameters $\theta$. For example, if we assumed each random variable was Poisson distributed then the probabilityi mass function is

$$P(X = x) = f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{9}$$

and then the loglikelihood is

$$\begin{aligned}
\ell\ell(\theta) &= \sum_{i=1}^{n} \log[\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}] && (10)\\
&= \sum_{i=1}^{n} \log\left(e^{-\lambda}\lambda^{x_i}\right) - \log(x_i!) && (11)\\
&= \sum_{i=1}^{n} \log\left(e^{-\lambda}\right) + \log(\lambda^{x_i}) - \log(x_i!) && (12)\\
&= \sum_{i=1}^{n} -\lambda + x_i \log(\lambda) - \log(x_i!) && (13)\\
& && (14)
\end{aligned}$$

Maximizing this function would return the maximum likelihood estimator for $\lambda$ or $\hat{\lambda}$. But $\hat{\lambda}$, and any other maximum likelihood esitmate, depends on our data. If we samples a different set of $n$ data points then we would find a different $\hat{\lambda}$.

Last week's notes showed a histogram of MLEs that looked suspicously "Normal"

### 1.0.3 Fundemental Theorum of Maximum likelihood estimates

The FTML says that the maximum likelihood esitmate is centered over the true parameter value (lets say $\lambda$) and normally distributed

$$\lambda \sim \mathcal{N}\left(\hat{\lambda}, \mathcal{I}^{-1}/N\right) \tag{15}$$

where $\mathcal{I}$ is the **Fisher Information** a function of the model's parameters that helps describe how they vary for repeated samples. It is often convienant to make clear that $\mathcal{I}$ is the Fisher information for a single data point $\mathcal{I}_1$ and to further define the Fisher information for $N$ data points as $\mathcal{I}_N = N\mathcal{I}$.

We can rewrite the FTML using $\mathcal{I}_N$ as

$$\lambda \sim \mathcal{N}\left(\hat{\lambda}, \mathcal{I}_N^{-1}\right) \tag{16}$$

For example, for the Poisson model above the Fisher Information is

$$\mathcal{I}_1 = \frac{1}{\lambda} \tag{17}$$

This means the Fisher information for $N$ data points is $N\mathcal{I}$ or

$$\mathcal{I}_N = N\mathcal{I}_1 = \frac{N}{\lambda} \tag{18}$$

However, we have a bit of a problem with the the Fisher information for $N$ datapoints. The $\mathcal{I}_N$ involves the **true** parameter value $\lambda$. We will need to estimate $\mathcal{I}_N$ and we can do that by replacing $\lambda$ with our estimate of $\lambda$ that we call $\hat{\lambda}$. The mle of $\mathcal{I}_N$ is then $\frac{N}{\hat{\lambda}}$. Note the small chance in notation but **big** change in meaning. We cannot compute the true Fisher information because we do not know the true parameter value $\lambda$. We can only estimate $\mathcal{I}_N$.

The MLE for the Poisson model, the value that maximizes the loglikelihood above, can be found analytically

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{19}$$

This means (finally) that we can chaarcterize the variability in our estimate or best guess for what the true $\lambda$ is:

$$\lambda \sim \mathcal{N}\left(\frac{\sum_{i=1}^{N} x_i}{N}, \left[\frac{N}{\hat{\lambda}}\right]^{-1}\right) \tag{20}$$

$$\lambda \sim \mathcal{N}\left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\hat{\lambda}}{N}\right) \tag{21}$$

$$\lambda \sim \mathcal{N}\left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\frac{1}{N}\sum_{i=1}^{N} x_i}{N}\right) \tag{22}$$

$$\lambda \sim \mathcal{N}\left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N} x_i}{N^2}\right) \tag{23}$$

$$\tag{24}$$

Lets put this theory to work.

Suppose we collect the following data:

$x = [38, 96, 24, 90, 60]$

and further we assume these datapoints were generated from the same Poisson distribution with parameter $\lambda$. Then we can say something about the probability of the **true,exact** $\lambda$.

$$\lambda \sim \mathcal{N}\left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N} x_i}{N^2}\right) \tag{25}$$

We can compute the sum of all xs

$$\sum_{i=1}^{N} x_i = 38 + 96 + 24 + 90 + 60 = 308 \tag{26}$$

and we know that $N = 5$ and so

$$\lambda \sim \mathcal{N}\left(308/5, 308/25\right) \tag{27}$$

$$\lambda \sim \mathcal{N}\left(61.6, 12.32\right) \tag{28}$$

$$\tag{29}$$

[31]:
```
import scipy

data = [38,96,24,90,60]
N = len(data)

fig,axs = plt.subplots(1,2)
```

```
ax=axs[0]
ax.scatter(data,[0]*N)

domain = np.arange(20,120)
probs = scipy.stats.poisson(61.6).pmf(domain) # most likely lambda

ax.plot(domain,probs,label="Estimated probability\n mass function")
ax.set_xlabel("Values of Xs")
ax.set_ylabel("Probability")


ax.legend(loc="upper right")

ax = axs[1]

domain = np.linspace(45,100,10**3)
model = scipy.stats.norm(61.6, np.sqrt(12.32) )

ax.plot(domain,model.pdf(domain))

ax.axvline(61.6,label="Maximum Likelihood Estimate",linestyle="--")

ax.legend()

ax.set_xlabel(r"Potential $\lambda$ values")
ax.set_ylabel("Density")

fig.set_size_inches(10,4)
fig.set_tight_layout(True)
```