MTH 224O, Spring 2024

Instructor: Bahman Angoshtari

Lecture 22

Sections 6.4: The mean squared error (MSE), the maximum likelihood estimation (MLE).

0. Overview of the lecture

- In this lecture, we discuss how to measure the quality of an estimator by the **mean squared error (MSE)**, and how to obtain good estimators using the **maximum likelihood estimator (MLE)**.

In [1]:
```
# Setting parameters of the Jupyter notebook
# This cell is only usefull if your are using Jupyter
sc = 0.75
options(repr.plot.width=16*sc,
        repr.plot.height=6*sc,
        repr.plot.pointsize = 20, # Text height in pt
        repr.plot.bg        = 'white',
        repr.plot.antialias = 'gray',
        #nice medium-res DPI
        repr.plot.res       = 300,
        #jpeg quality bumped from default
        repr.plot.quality   = 90,
        #vector font family
        repr.plot.family    = 'serif', # Vector font family. 'sans', 'serif'
        "getSymbols.warning4.0"=FALSE)
```

1. The mean squared error (MSE)

- **MSE** is a common criterion to evaluate an estimator.
- Let $\hat{\theta}$ be an estimator for a parameter θ. Note that $\hat{\theta}$ is a random variable and that θ is an unknown number.
- The error in estimating θ with $\hat{\theta}$ is $\epsilon = \theta - \hat{\theta}$.

- The mean squared error of $\hat{\theta}$ is: MSE$_{\hat{\theta}} = \mathbb{E}[\epsilon^2] = \mathbb{E}\left[(\theta - \hat{\theta})^2\right]$
- A good estimator should have a small MSE.

For any estimator $\hat{\theta}$ of a parameter θ, we define two numbers:

- Bias of $\hat{\theta} = \theta - \mathbb{E}[\hat{\theta}]$.
  - Bias measures *on average* how wrong an estimator is.
- Variance of $\hat{\theta} = \mathrm{Var}(\hat{\theta})$.
  - Variance of an estimator measures how variable the estimator is.
- There is a close relationship between MSE, bias, and variance.

Theorem: For any estimator $\hat{\theta}$ of a parameter θ we have:

$$\mathrm{MSE}_{\hat{\theta}} = (\mathrm{Bias}_{\hat{\theta}})^2 + \mathrm{Var}(\hat{\theta})$$

Proof:

$$\begin{aligned}
\mathrm{MSE}_{\theta} &= \mathbb{E}\left[(\theta - \hat{\theta})^2\right] \\
&= \mathbb{E}\left[(\theta - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \hat{\theta})^2\right] \\
&= \mathbb{E}\left[(\theta - \mathbb{E}[\hat{\theta}])^2 + (\mathbb{E}[\hat{\theta}] - \hat{\theta})^2 + 2(\theta - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \hat{\theta})\right] \\
&= (\theta - \mathbb{E}[\hat{\theta}])^2 + \mathbb{E}\left[(\mathbb{E}[\hat{\theta}] - \hat{\theta})^2\right] + 2(\theta - \mathbb{E}[\hat{\theta}])\mathbb{E}\left[\mathbb{E}[\hat{\theta}] - \hat{\theta}\right] \\
&= (\mathrm{Bias}_{\hat{\theta}})^2 + \mathrm{Var}(\hat{\theta}) + 0.
\end{aligned}$$

Example 1:

- Let $\{X_n\}_{n=1}^{N} \overset{\text{i.i.d.}}{\sim}$ Bern(p).
- Define $\hat{p} := \overline{X} = \frac{X_1 + \cdots + X_N}{N}$ to be the sample mean.
- Find the MSE$_{\hat{p}}$ (for estimating p).

Solution:

- By the theorem above, we have that $\mathrm{MSE}_{\hat{p}} = (\mathrm{Bias}_{\hat{p}})^2 + \mathrm{Var}(\hat{p})$.

- Since $\sum_{n=1}^{N} X_n = N\hat{p}$ $\sim$Binomial(N, p), we have $\mathbb{E}[\hat{p}] = \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N} X_n\right] = \frac{1}{N}Np = p$.

  So, $\mathrm{Bias}_{\hat{p}} = p - \mathbb{E}[\hat{p}] = 0$. We say that $\hat{p}$ is **unbiased**.

- Similarly, Var$(\hat{p}) = \frac{1}{N^2}\mathrm{Var}\left(\sum_{n=1}^{N} X_n\right) = \frac{1}{N^2}Np(1-p) = \frac{p(1-p)}{N}$.

- Finally, we obtain that MSE$_{\hat{p}} = (\mathrm{Bias}_{\hat{p}})^2 + \mathrm{Var}(\hat{p}) = 0 + \frac{p(1-p)}{N} = \frac{p(1-p)}{N}$.

2. Maximum Likelihood Estimation (MLE)

- How can we find a good estimator?
- Generally, we consider a criterion for how good an estimator is, and then try to find an estimator that optimize that criterion.
- The method of **maximum likelihood estimation** uses the **likelihood** function as the criterion.

Definition: Let $X_1, \ldots, X_N$ be an i.i.d. sample.

- If X_n is discrete with p.m.f. $p(x;\theta)$, then the likelihood function is

$$L(\theta) = \prod_{n=1}^{N} p(X_n;\theta) = p(X_1;\theta) \times p(X_2;\theta) \times \cdots \times p(X_N;\theta)$$

- If X_n is continuous with p.d.f. $f(x;\theta)$, then the likelihood function is

$$L(\theta) = \prod_{n=1}^{N} f(X_n;\theta) = f(X_1;\theta) \times f(X_2;\theta) \times \cdots \times f(X_N;\theta)$$

- What is the meaning of the likelihood function?

- The likelihood function $L(\theta)$ is a measure of the "likelihood" of observing the values $X_1, \ldots, X_N$ if the actual value of the parameter is θ

- This can be easily checked for the discrete case. Indeed, if X_n has marginal p.m.f. $p(x_n; \theta)$, then

$$\mathbb{P}(X_1 = x_1 \cap \cdots \cap X_N = x_n) = p(x_1; \theta) \times \cdots \times p(x_N; \theta)$$

- So, $L(\theta)$ is indeed the probability of getting the observed values $X_1, \ldots, X_N$ if the actual value of the parameter is θ.

- For the continuous case, the idea is similar. However, one should use the joint pdf, which is outside the scope of this course.

Example 2: Let $\{X_n\}_{n=1}^N \overset{\text{i.i.d.}}{\sim}$ Pois(λ). Find the likelihood function $L(\lambda)$.

Solution:

- Since pmf of X_n is $p(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, the likelihood function is

$$\begin{aligned} L(\lambda) &= p(X_1; \lambda) \times p(X_2; \lambda) \times \cdots \times p(X_N; \lambda) \\ &= \left(\frac{\lambda^{X_1}}{X_1!} e^{-\lambda} \right) \times \left(\frac{\lambda^{X_2}}{X_2!} e^{-\lambda} \right) \times \cdots \times \left(\frac{\lambda^{X_N}}{X_N!} e^{-\lambda} \right) \\ &= \frac{1}{X_1! X_2! \ldots X_N!} \lambda^{X_1 + X_2 + \cdots + X_N} e^{-N\lambda}. \end{aligned}$$

- Let us do a little bit of experiment.

In [2]:
```
X = rpois(10, lambda=3) # generating a random Poisson sample
X
```

1 · 2 · 0 · 2 · 6 · 4 · 5 · 3 · 4 · 1

In [3]:
```
3^5/factorial(5)*exp(-3)
```

0.100818813444924

In [4]:
```
dpois(5,lambda=3)  # the pmf of Poisson
```

0.100818813444924

In [5]:
```
dpois(X,lambda=3)
```

0.149361205103592 · 0.224041807655388 · 0.0497870683678639 · 0.224041807655388 · 0.0504094067224622 · 0.168031355741541 · 0.100818813444924 · 0.224041807655388 · 0.168031355741541 · 0.149361205103592

In [6]:
```
prod(dpois(X,lambda=3)) # the likelihood function
```

1.79231032826978e-09

In [7]:
```
prod(dpois(X,lambda=10))
```

3.11461479366671e-25

In [8]:
```
L = function(lam) prod(dpois(X,lambda=lam))
L(3)
L(10)
sapply(c(3,10), L)
```

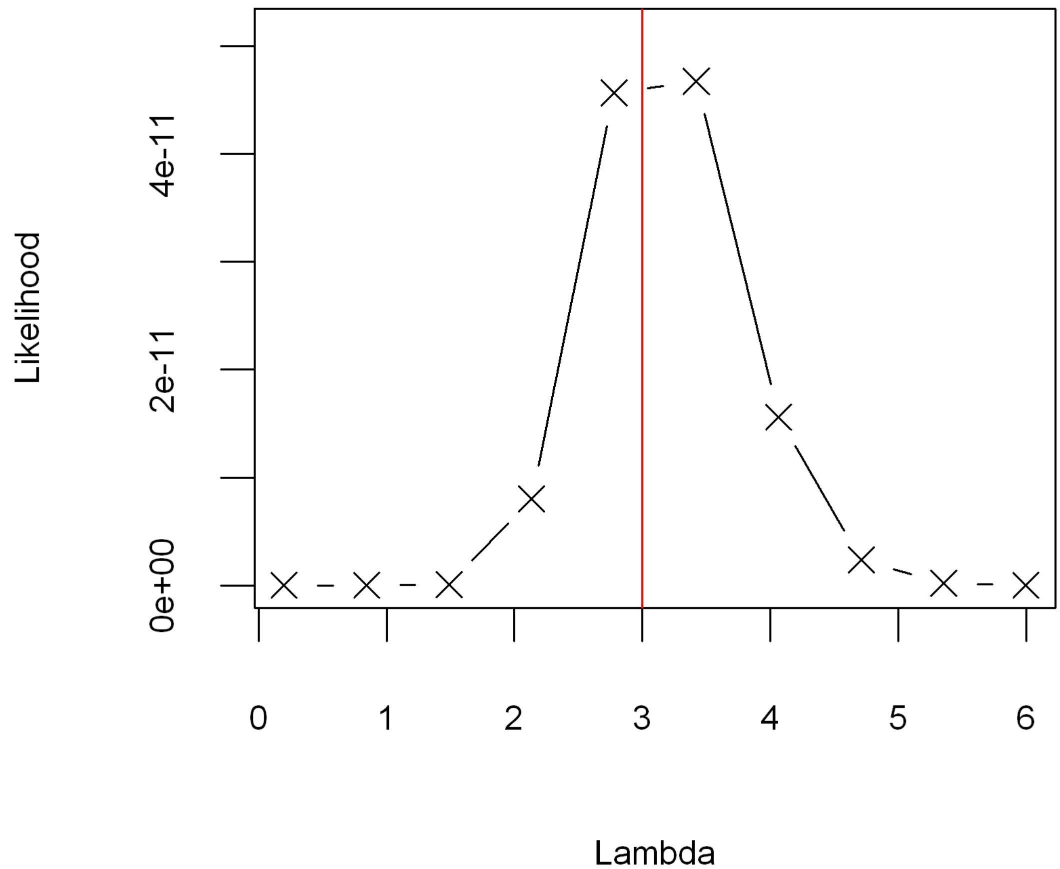
1.79231032826978e-09

3.11461479366671e-25

1.79231032826978e-09 · 3.11461479366671e-25

In [9]:
```
lam=3
N = 10
X = rpois(N, lambda=lam) # generating a random Poisson sample
L = function(lam) prod(dpois(X,lambda=lam))
lambda_values = seq(0.2,6,length.out=10)
Likelihood_values = sapply(lambda_values, L)
options(repr.plot.width=6, repr.plot.height=6)
max(Likelihood_values)
plot(
    lambda_values, Likelihood_values,
    type='b', pch=4, ylim=c(0,1.1*max(Likelihood_values)),
    xlab="Lambda", ylab="Likelihood"
)
abline(v=lam, col='red')
```

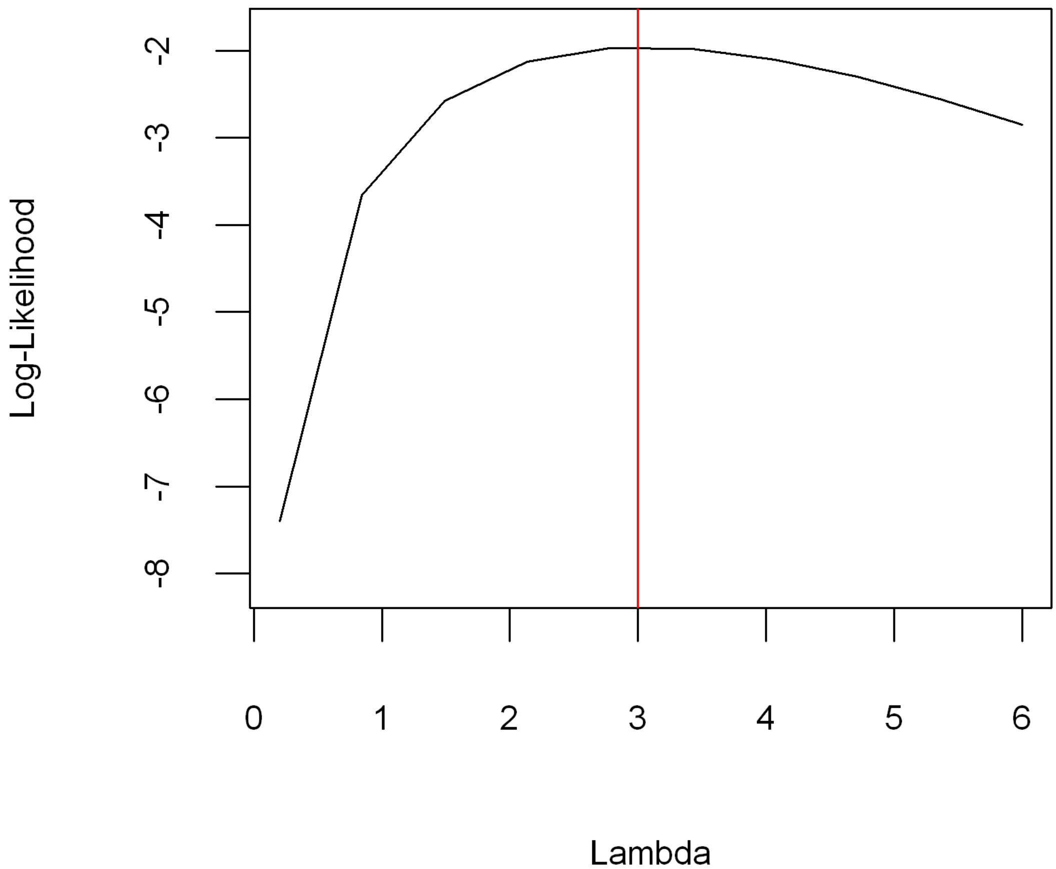
4.6734991954761e-11



In [10]:

```
lam=3
N = 3000
X = rpois(N, lambda=lam) # generating a random Poisson sample
LogL = function(lam) sum(log(dpois(X,lambda=lam)))/N
lambda_values = seq(0.2,6,length.out=10)
LogLikelihood_values = sapply(lambda_values, LogL)
options(repr.plot.width=6, repr.plot.height=6)
max(LogLikelihood_values)
plot(
    lambda_values, LogLikelihood_values,
    type='l', pch=4, ylim=c(1.1*min(LogLikelihood_values),0.9*max(LogLikelih
    xlab="Lambda", ylab="Log-Likelihood"
)
abline(v=lam, col='red')
```

-1.9723671389017



- Since larger values of the likelihood function are better, we maximize the likelihood function $L(\theta)$ to obtain a good estimator for the parameter θ.
- The resulting estimator is called the “maximum likelihood estimator (MLE)” of θ.

$$\hat{\theta}_{\mathrm{MLE}} = \max_{\theta} L(\theta)$$

- The reason for popularity of the maximum likelihood estimators it that, for most pdf and pmf:
  - $\hat{\theta}_{\mathrm{MLE}}$ is asymptotically unbiased: $\lim_{N\to+\infty} \mathbb{E}[\hat{\theta}_{\mathrm{MLE}}] = \theta$.
  - $\hat{\theta}_{\mathrm{MLE}}$ is asymptotically the minimum variance estimator:

$$\lim_{N\to+\infty} \mathrm{Var}(\hat{\theta}_{\mathrm{MLE}}) = \min\left\{\mathrm{Var}(Y) : Y \text{ is an estimator of } \theta\right\}.$$

Example 3: Let $\{X_n\}_{n=1}^{N} \overset{\text{i.i.d.}}{\sim}$ Pois(λ). Find the MLE of λ.

Solution:

- As we calculated in Example 2 above, the likelihood function is

$$L(\lambda) = \frac{1}{X_1!X_2!\dots X_N!}\lambda^{X_1+X_2+\cdots+X_N}e^{-N\lambda}.$$

- To maximize $L(\lambda)$, we should differentiate it and set equal to zero (to find its stationary point). The derivative $L'(\lambda)$ may however be a bit complicated to work with (try differentiating, it is not too bad).

- In most cases, it would be easier to maximize the **log-likelihood** function $\mathcal{L}(\lambda) = \log(L(\lambda))$. Note that the value of λ that maximizes $L(\lambda)$ is the same as the value of λ that maximizes $\mathcal{L}(\lambda)$ (why?).

- The log-likelihood function is

$$\begin{aligned}\mathcal{L}(\lambda) = \log(L(\lambda)) &= \log\left(\frac{1}{X_1!X_2!\dots X_N!}\lambda^{X_1+X_2+\cdots+X_N}e^{-N\lambda}\right)\\ &= -\log(X_1!X_2!\dots X_N!) + (X_1+X_2+\cdots+X_N)\log(\lambda) - N\lambda.\end{aligned}$$

- The MLE is the maximizer of the log-likelihood function. To find it, we calculate as follows:

$$\mathcal{L}(\lambda) = -\log(X_1!X_2!\dots X_N!) + (X_1+X_2+\cdots+X_N)\log(\lambda) - N\lambda.$$

$$\begin{aligned}\mathcal{L}'(\hat{\lambda}) &= \frac{X_1+X_2+\cdots+X_N}{\hat{\lambda}} - N = 0\\ \implies \qquad \hat{\lambda} &= \frac{X_1+X_2+\cdots+X_N}{N} = \overline{X}.\end{aligned}$$

- Since $\mathcal{L}''(\lambda) = -\frac{X_1+X_2+\cdots+X_N}{\lambda^2} < 0$, the above $\hat{\lambda}$ is the unique maximizer of $\mathcal{L}(\lambda)$ (and $L(\lambda)$). So, it is the MLE for λ.

- In other words, the MLE of λ is the sample mean. This makes sense, since λ is the mean of the Pois(λ) distribution.

In [ ]:

```

```