LECTURE 3

Let us recall a few key results from Lecture 2 that will be used here:

* $[X_1\ X_2 \dots X_n]^T$ random sample generated by sampling the RV $X \sim (\mu_X, \sigma_X^2)$; $(X_1, X_2, \dots, X_n)$ - i.i.d. $\Rightarrow$ $\hat{X} \triangleq \frac{1}{n}\sum X_i$, $E_{\hat{X}}(\hat{X}) = \mu_X$, $E_{\hat{X}}\left((\hat{X}-\mu_X)^2\right) = \sigma_X^2/n$ □

* $(X_1, X_2, \dots, X_n)$ converges in distribution to X $\overset{\text{DEF}}{\Leftrightarrow}$ $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)\ \ \forall x \in \mathbb{R}$, where $F_X(\cdot)$ and $F_{X_n}(\cdot)$ are cdfs □

* $(X_1, X_2, \dots, X_n)$ converges in probability to a constant c^* $\overset{\text{DEF}}{\Leftrightarrow}$ $(X_1, X_2, \dots, X_n)$ converges in distribution to a RV X such that: $P(X = c^*) = 1$ □

* LLN: $(X_1, X_2, \dots, X_n)$ i.i.d. $\Rightarrow$ Denoted with $\hat{X}_i \triangleq \frac{1}{i}\sum_{j=1}^{i} X_j$, we have that $\{\hat{X}_i\}_i \xrightarrow{P} \mu_X$ □

* CLT: $(X_1, X_2, \dots, X_n)$ i.i.d. $\Rightarrow$ Denoted with $Z_i \triangleq \frac{\sqrt{i}}{\sigma_X}(\hat{X}_i - \mu_X)$, we have that $\{Z_i\}_i \xrightarrow{D} N(0,1)$ □

* Finally, let us recall that, given two RVs X and Y with joint probability function $f(x,y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$, respectively, we can write:

$$f_{X|Y}(x|y) \triangleq f_X(x|Y=y) = \frac{f(x,y)}{f_Y(y)}$$ - Conditional pdf of X given $Y=y$

A generalization of the Bayes' theorem can be obtained by noticing that:

- $f(x,y) = f_{X|Y}(x|y)\, f_Y(y) = f_{Y|X}(y|x)\, f_X(x)$

- $f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx = \int_{-\infty}^{+\infty} f_{Y|X}(y|x)\,f_X(x)\,dx$

Bayes' Theorem $\quad f_{X|Y}(x|y) = \dfrac{f_{Y|X}(y|x)\,f_X(x)}{\int_{-\infty}^{+\infty} f_{Y|X}(y|x)\,f_X(x)\,dx}$ (if X is continuous)

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\,f_X(x)}{\sum_x f_{Y|X}(y|x)\,f_X(x)} \quad \text{(if X is discrete)}$$

□

A few well-studied probability functions will be considered here:

- Bernoulli RV : $X \sim \text{Bernoulli}(p) \overset{\text{DEF}}{\iff} P(X=1)=p\ ;\ P(X=0)=1-p$

  For this RV, we have: $\mu_X = p\ ;\ \sigma_X^2 = p(1-p)$

- Binomial RV: $X \sim B(n,p) \overset{\text{DEF}}{\iff} P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k=0,1,\dots,n$

  For this RV, we have: $\mu_X = np\ ;\ \sigma_X^2 = np(1-p)$

- Poisson RV: $X \sim P(\lambda) \overset{\text{DEF}}{\iff} P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad k=0,1,2,\dots$

  For this RV, we have: $\mu_X = \lambda\ ;\ \sigma_X^2 = \lambda$. This RV is a generalization of $B(n,p)$ as $p \to 0$ and $n \to \infty$ with $\lambda = np$

Note that in all the RVs considered above, we assume that the binary events they are built upon are INDEPENDENT

□

Let us consider the following scenario: we have a random sample $(X_1, X_2, \dots, X_n)$ generated by a partially-unknown RV X and we would like to estimate X given the random sample.

- We know about X: $X \sim f_X(x) = f_X(x|\underline{\theta})$ with f_X belonging to a known class of pdfs
  ($\underline{\theta}$: parameter vector)
- We do NOT know about X: θ^* (i.e., the true value of θ)
- We would like to: $(x_1, x_2, \dots, x_n) \longrightarrow$ Estimation $\hat{\theta}$ (*)
  where $X_1 = x_1$, $X_2 = x_2$, ...

Condition (*), though, means that the estimation $\hat{\theta}$ may vary when the n-tupla $(x_1, x_2, \dots, x_n)$ changes $\Rightarrow$ We can define a new RV:

Estimator: $T: (x_1, x_2, \dots, x_n) \longrightarrow \hat{\theta}$

$\Downarrow$

$T = T(X_1, X_2, \dots, X_n)$ - is a function of RVs and its own distribution must be determined

The definition of T (and, hence, its distribution) depends on how $\hat{\theta}$ is estimated by using $(x_1, x_2, \dots, x_n)$, i.e., it depends on the method used to fit $f_x(\cdot)$ on data. The method of Maximum Likelihood (ML) choses $\hat{\theta}$ such that:

$$\hat{\theta} = \arg\max_{\theta} f_X(x|\theta)$$

for every observed value x

Ex.: $X \sim B(n, p)$ with p-unknown, i.e., $\theta \triangleq p$. In this case, we have:

$$f_x(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad x = 0, 1, 2, \dots, n \text{ and } f_X(x) \triangleq P(X = x)$$

In this case, it is easy to show that the value $\hat{\vartheta}$ satisfies:

$$\frac{\partial f_x}{\partial \vartheta} = 0 \iff \frac{\partial}{\partial \vartheta}\left[\vartheta^x (1-\vartheta)^{n-\vartheta}\right] = 0$$

We can drop the constant $\binom{n}{x}$

$$\iff x\vartheta^{x-1}(1-\vartheta)^{n-x} - (n-x)\vartheta^x(1-\vartheta)^{n-x-1} = 0$$

$$\iff \vartheta^{x-1}(1-\vartheta)^{n-x-1}\left[x(1-\vartheta) - (n-x)\vartheta\right] = 0$$

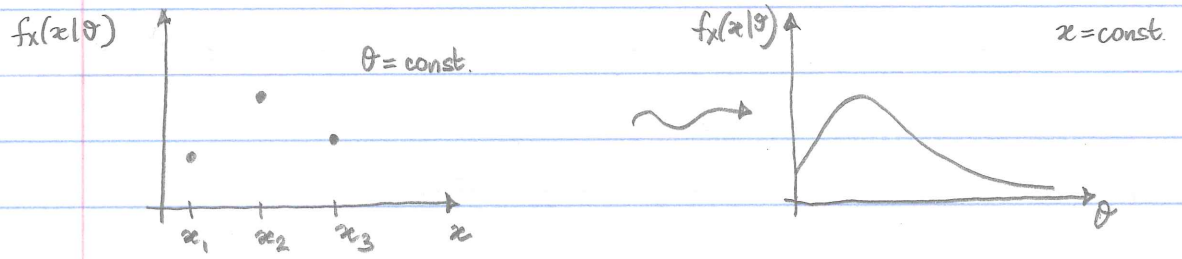
$$\iff \vartheta^{x-1}(1-\vartheta)^{n-x-1}\left[x - \vartheta n\right] = 0$$

$$\iff \hat{\vartheta} = \frac{x}{n} \qquad (**)$$

Formula $(**)$ shows: $T_n : x \to \hat{\vartheta} = \frac{x}{n}$ is a RV. In fact: $T_n = \frac{X}{n}$ and, by the LLN, we have: $\frac{X}{n} \xrightarrow{P} \vartheta^*$, i.e., the true value of the probability that a Bernoulli trial results in 1. $\Rightarrow$ T_n is called the MAXIMUM LIKELIHOOD ESTIMATOR (MLE)

This example highlights a few facts:

- With the ML method, we use $f_X(x|\vartheta)$ as a function of ϑ, i.e., we fix x and let ϑ vary $\Rightarrow$ The maximization problem is now less affected by the amount of data (i.e., the number of samples $x_1, x_2, \dots, x_n$) we have



- The condition $\max_{\vartheta} f_X(x|\vartheta)$ is defined up to a positive constant
  $\Rightarrow$ We can choose a function $L(\vartheta)$ to be maximized such that $L(\vartheta) \propto f_X(x|\vartheta)$
  $\Rightarrow$ $L(\vartheta)$ is the LIKELIHOOD FUNCTION

In our example, $L(\vartheta) \triangleq \vartheta^x (1-\vartheta)^{n-x}$

- The ML method can be generalized to a random sample, i.e., we replace $f_x(\cdot)$ with the joint probability function $f(x_1, x_2, \dots, x_n)$ and choose:

$$L(\vartheta) \propto f(x_1, x_2, \dots, x_n | \vartheta)$$

- In many cases, it may be appropriate to maximize the LOG-LIKELIHOOD function $l(\vartheta) \triangleq \log(L(\vartheta))$ instead of the likelihood function. In our example:

$$L(\vartheta) = \vartheta^x (1-\vartheta)^{n-x} \Rightarrow l(\vartheta) = x \log\vartheta + (n-x)\log(1-\vartheta)$$

Hence: $\frac{\partial L}{\partial \vartheta} = 0 \Leftrightarrow \frac{\partial l}{\partial \vartheta} = 0 \Leftrightarrow \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta} = 0 \Leftrightarrow \hat{\vartheta} = \frac{x}{n}$

Ex.: Let us consider n samples $x_1, x_2, \dots, x_n$ from the RV $X \sim N(\mu, \sigma^2)$ where $\vartheta \triangleq \mu$ must be estimated and σ^2 is known. $\Rightarrow$ We can think $x_1, x_2, \dots, x_n$ as the outcome of a random sample $(X_1, X_2, \dots, X_n)$ where $X_i \sim N(\mu, \sigma^2)$ $\forall i$ and the RVs are independent $\Rightarrow$ We have:

$$f(x_1, x_2, \dots, x_n | \vartheta) = \prod_{i=1}^{n} f_{x_i}(x_i|\vartheta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\vartheta)^2}{\sigma^2}}$$

We define: $l(\vartheta) \triangleq \log\left(\prod_{i=1}^{n} f_{x_i}(x_i|\vartheta)\right) = -\sum_{i=1}^{n} \frac{(x_i-\vartheta)^2}{\sigma^2} =$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\vartheta + \vartheta^2\right) = -\frac{n}{\sigma^2}\vartheta^2 + \frac{2}{\sigma^2} n\bar{x}\,\vartheta - \overbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^2}^{\text{It is constant}}$$

$\bar{x}$: sample mean

Hence: $\frac{\partial \ell}{\partial \vartheta} = 0 \iff -2\vartheta \frac{n}{\sigma^2} + 2\frac{n}{\sigma^2}\bar{x} = 0 \iff \hat{\vartheta} = \bar{x}$

In this example we have that the MLE $T_n = \hat{X}_n$ (sample mean) and, because of the LLN, we have: $T_n \xrightarrow{P} \mu$.

- The ML estimation is an optimization problem:

(***) $\max \ell(\vartheta)$

s.t.: $\vartheta \in C$ with C set of admissable parameters

As in any optimization problem, ϑ can be either a scalar or a vector and the existence (and uniqueness) of the solution depends on the definition of $\ell(\vartheta)$ and C. In particular, let us note:

- The class of the function $f_X(\cdot)$ determines whether $\ell(\vartheta)$ is concave or not and whether it has one maximum or many local maxima

⇓

We will use generalized linear models and Gaussian mixtures

- If $\ell(\vartheta)$ is a known mathematical function and concave ⇒ The solution exists and can be computed by solving: $\left.\frac{\partial \ell}{\partial \vartheta}\right|_{\vartheta=\hat{\vartheta}} = 0$

- If $\ell(\vartheta)$ is concave and numerically evaluated ⇒ We can apply algorithms for global maximization (e.g. Newton's method, quasi-Newton's methods, secant method, etc.) ⇒ These methods are invoked by MLE routines in MATLAB, R, etc.

- From a numerical standpoint, the optimization problem is often formulated as a minimization problem $\Rightarrow$ We solve for $\hat{\vartheta} = \arg\min_{\vartheta \in \Theta} \left(-\ell(\vartheta)\right)$

Because the MLE T_n is a RV, we must characterize the probability distribution of $T_n \Rightarrow$ The features of T_n as RV will justify its wide use as an estimation technique:

* ML Estimators are consistent and asymptotically normal

T_n - consistent $\overset{DEF}{\Longleftrightarrow}$ $T_n \xrightarrow{P} \vartheta^*$ (true set of parameters), i.e.,

$$\forall \varepsilon > 0 \quad P\left(|T_n - \vartheta^*| > \varepsilon\right) \longrightarrow 0 \quad \text{as } n \longrightarrow +\infty$$

T_n - asymptotically normal $\overset{DEF}{\Longleftrightarrow}$ $\dfrac{T_n - \vartheta^*}{\sigma_{T_n}} \xrightarrow{D} N(0,1)$ where $\sigma_{T_n} \triangleq$ standard dev. of T_n

The consistency of the MLE can be shown by noticing that MLE minimizes the mean square error (MSE). In fact, denoted with $MSE(T_n) \triangleq E_{T_n}\left((T_n - \vartheta^*)^2\right)$, the following inequality holds:

$$\forall \varepsilon > 0 \quad P\left(|T_n - \vartheta^*| > \varepsilon\right) < \frac{E_{T_n}\left((T_n - \vartheta^*)^2\right)}{\varepsilon^2}$$

Hence, the condition $MSE(T_n) \to 0$ shall imply that $T_n \xrightarrow{P} \vartheta^*$. It is interesting that MLE minimizes MSE because the following occurs:

$$(T_n - \vartheta^*)^2 = (T_n - \underbrace{\mu_n}_{\triangleq E_{T_n}(T_n)})^2 + (\mu_n - \vartheta^*)^2 + 2(T_n - \mu_n)(\mu_n - \vartheta^*) \Rightarrow$$

(expected value)

By applying the expected value operator and exploiting the linearity:

$$MSE(T_n) = \underset{T_n}{E}\left((T_n-\mu_n)^2\right) + (\mu_n-\vartheta^*)^2 + 2(\mu_n-\vartheta^*)\underbrace{\underset{T_n}{E}(T_n-\mu_n)}_{0} \Rightarrow$$

($\underset{T_n}{E}((T_n-\mu_n)^2)$: variance of T_n; $(\mu_n-\vartheta^*)^2$: a constant)

$$MSE(T_n) = \underset{T_n}{E}\left((T_n-\mu_n)^2\right) + \left(\underset{T_n}{E}(T_n-\vartheta^*)\right)^2 \Rightarrow$$

To minimize MSE is equivalent to minimize the bias (squared) plus the variance of T_n

Ex.: Let us suppose that a train of neural spikes follow a Poisson distribution with unknown parameter λ (i.e., $\vartheta^* = \lambda$) and let us assume that we have n independent spike counts:

$$X \sim \mathcal{P}(\lambda) \Rightarrow f_X(x) \triangleq P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f_X(x_i) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Let us define: $\ell(\vartheta) = \log\left(e^{-n\vartheta}\vartheta^{\sum_i x_i}\right) = -n\vartheta + \log\vartheta \sum_{i=1}^{n} x_i$

By applying the ML method, we have: $\frac{\partial \ell}{\partial \vartheta} = 0 \Leftrightarrow -n + \frac{1}{\vartheta}\sum_{i=1}^{n} x_i = 0$

$$\Leftrightarrow \hat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$, i.e., the MLE is the sample mean $\bar{X}$

In this case, we already know that $\underset{T_n}{E}(T_n) = \underset{\bar{X}}{E}(\bar{X}) = \lambda \Rightarrow MSE(T_n) = \underset{T_n}{E}\left((T_n-\lambda)^2\right)$ i.e., the MLE is unbiased

Now, in order to show that $MSE(T_n)$ is less than the MSE obtained with other estimators, let us consider the estimator:

$$S_n^2 \triangleq \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \quad \text{(sample variance)}$$

We have: $E\left(S_n^2\right) = \frac{1}{n-1}\sum_{i=1}^{n} E\left((X_i-\bar{X})^2\right) =$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E\left((X_i-\lambda)^2+(\lambda-\bar{X})^2-2(X_i-\lambda)(\lambda-\bar{X})\right)$$

($\mu_X=\lambda$)

$$= \frac{1}{n-1}\left[n\sigma_X^2 + n\cdot\frac{\sigma_X^2}{n} - 2n\frac{\sigma_X^2}{n}\right] = \sigma_X^2 = \lambda$$

($\uparrow$ Poisson)

Hence: $MSE\left(S_n^2\right) = \underset{S_n^2}{E}\left((S_n^2-\lambda)^2\right)$ - unbiased

Now, by computing the variances, one obtains:

$MSE(T_n) = \frac{\lambda}{n}$ (see lecture 2, page 13)

$MSE(S_n^2) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1}$

$\Rightarrow MSE(T_n) < MSE(S_n^2) \quad \forall n>0$

□

With regard to the asymptotic normality, one can easily prove it in this case:

$$\max_{\vartheta\in\Theta} \ell(\vartheta) \Longleftrightarrow \left.\frac{\partial \ell}{\partial\vartheta}\right|_{\vartheta=\hat{\vartheta}} = 0$$

Let us assume: $\ell(\vartheta) = \frac{1}{n}\log f(x_1,x_2,\dots,x_n|\vartheta) \overset{\text{i.i.d. assumption}}{=} \frac{1}{n}\sum_{i=1}^{n}\log f_X(x_i|\vartheta)$

So we can write: $\frac{\partial \ell}{\partial\vartheta}(\hat{\vartheta}) = 0 \Longleftrightarrow \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\left(\log f_X(x_i|\vartheta)\right)_{\vartheta=\hat{\vartheta}} = 0$

Let us remind: $\frac{\partial}{\partial\vartheta}\left(\log f_X(X|\vartheta)\right) = \frac{f_X'(X|\vartheta)}{f_X(X|\vartheta)}$ where $f_X'(X|\vartheta) \triangleq \frac{\partial f_X}{\partial\vartheta}(X|\vartheta)$

$$\frac{\partial^2}{\partial\vartheta^2}\left(\log f_X(X|\vartheta)\right) = \frac{f''(X|\vartheta)}{f_X(X|\vartheta)} - \frac{\left[f'(X|\vartheta)\right]^2}{f_X^2(X|\vartheta)} \quad \text{where } f_X''(X|\vartheta) \triangleq \frac{\partial^2 f_X}{\partial\vartheta}(X|\vartheta)$$

Therefore: $\underset{X}{E}\left(\frac{\partial}{\partial\vartheta}\log f_X(X|\vartheta)\right) = \int_{-\infty}^{+\infty} f'_X(x|\vartheta)\,dx = \frac{\partial}{\partial\vartheta}\overbrace{\int_{-\infty}^{+\infty} f_X(x|\vartheta)\,dx}^{=1} = 0$

$$\underset{X}{E}\left(\frac{\partial^2}{\partial\vartheta^2}\log f_X(X|\vartheta)\right) = \underbrace{\int_{-\infty}^{+\infty} f''_X(x|\vartheta)\,dx}_{=0} - \int_{-\infty}^{+\infty}\frac{[f'_X(x|\vartheta)]^2}{f_X^2(x|\vartheta)} f_X(x|\vartheta)\,dx$$

$$= -\underset{X}{E}\left(\left(\frac{\partial}{\partial\vartheta}\log f_X(X|\vartheta)\right)^2\right)$$

The quantity $\underset{F}{I}(\vartheta) \triangleq -\underset{X}{E}\left(\frac{\partial^2}{\partial\vartheta^2}\log f_X(X|\vartheta)\right) = \underset{X}{E}\left(\left(\frac{\partial}{\partial\vartheta}\log f_X(X|\vartheta)\right)^2\right)$ is called

"Fisher Information" of the RV X

We can apply the mean value theorem (MVT) to $\partial\ell/\partial\vartheta$ between ϑ^* (i.e, the true parameter) and $\hat{\vartheta}$ (i.e, the ML estimation):

MVT: $\frac{f(a)-f(b)}{a-b} = f'(c) \quad c\in[a,b] \Rightarrow \quad a \triangleq \hat{\vartheta} \quad b \triangleq \vartheta^* \quad f(\vartheta) \triangleq \frac{\partial\ell}{\partial\vartheta}$

$$\frac{\partial\ell}{\partial\vartheta}(\hat{\vartheta}) = \frac{\partial\ell}{\partial\vartheta}(\vartheta^*) + \frac{\partial^2\ell}{\partial\vartheta^2}(\tilde{\vartheta})(\hat{\vartheta}-\vartheta^*) \quad \text{with } \tilde{\vartheta}\in[\hat{\vartheta},\vartheta^*] \text{ (assuming } \hat{\vartheta}<\vartheta^*)$$

$$\Rightarrow \sqrt{n}\,(\hat{\vartheta}-\vartheta^*) = -\frac{\sqrt{n}\,\frac{\partial\ell}{\partial\vartheta}(\vartheta^*)}{\frac{\partial^2\ell}{\partial\vartheta^2}(\tilde{\vartheta})}$$

Let us note: $\frac{\partial\ell}{\partial\vartheta}(\vartheta^*) \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\vartheta}\left(\log f_X(x_i|\vartheta)\right)_{\vartheta=\vartheta^*}$ - If we define:

$$Y_i \triangleq \frac{\partial}{\partial\vartheta}\left(\log f_X(X_i|\vartheta)\right)_{\vartheta=\vartheta^*} \Rightarrow \frac{\partial\ell}{\partial\vartheta}(\vartheta^*) = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ - sample mean}$$

Hence: - By LLN: $E\left(\frac{\partial\ell}{\partial\vartheta}(\vartheta^*)\right) = \underset{X}{E}\left(\frac{\partial}{\partial\vartheta}\log f_X(X|\vartheta)\Big|_{\vartheta=\vartheta^*}\right) = 0$

- By CLT: $\sqrt{n}\,\frac{\partial \ell}{\partial \vartheta}(\vartheta^*) \xrightarrow{D} N\left(0,\ var\left(\frac{\partial}{\partial \vartheta}\log f_X(X|\vartheta)\Big|_{\vartheta=\vartheta^*}\right)\right) =$

$$= N\left(0, I_F(\vartheta^*)\right) \quad (1)$$

Let us also note: $\frac{\partial^2 \ell}{\partial \vartheta^2}(\tilde{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial \vartheta^2}\left(\log f_X(x_i|\vartheta)\right)_{\vartheta=\tilde{\vartheta}}$ - If we define:

$$Z_i \triangleq \frac{\partial^2}{\partial \vartheta^2}\left(\log f_X(X_i|\vartheta)\right)_{\vartheta=\tilde{\vartheta}} \Rightarrow \frac{\partial^2 \ell}{\partial \vartheta^2}(\tilde{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n} Z_i \text{ - sample mean}$$

Hence, by LLN: $E\left(\frac{\partial^2 \ell}{\partial \vartheta^2}(\tilde{\vartheta})\right) = -I_F(\tilde{\vartheta})$

Also, since $T_n = \hat{\vartheta}$ is a consistent estimator, we have:

$$T_n \xrightarrow{P} \vartheta^* \Rightarrow I_F(\tilde{\vartheta}) \xrightarrow{P} I_F(\vartheta^*) \quad (2)$$

By combining (1) and (2) we have: $\sqrt{n}\left(T_n - \vartheta^*\right) \xrightarrow{D} N\left(0, \frac{1}{I_F(\vartheta^*)}\right)$

Finally one can note:

$$I_n(T_n) \triangleq -\sum_{i=1}^{n}\frac{\partial^2}{\partial \vartheta^2}\left(\log f_X(X_i|\vartheta)\right)_{\vartheta=\hat{\vartheta}} = n\, I_F(\hat{\vartheta})$$

$$T_n \xrightarrow{P} \vartheta^* \Rightarrow I_F(\hat{\vartheta}) \xrightarrow{P} I_F(\vartheta^*)$$

Slutsky's Theorem $\Rightarrow \sqrt{I_n(T_n)}\,(T_n - \vartheta^*) \xrightarrow{D} N(0,1)$

In this derivation, we have $\sigma^2_{T_n} = \frac{1}{I_n(T_n)}$. It can be proved that the variance of MLE is actually $1/I_n(T_n)$. To give a sense of this, let us consider $f_X(x|\vartheta) \triangleq \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\vartheta)^2}{2\sigma^2}}$ where σ is known. In this case, we have:

$$\ell(\vartheta) = \log e^{-\frac{(x-\vartheta)^2}{2\sigma^2}} \stackrel{c}{=} -\frac{(x-\vartheta)^2}{2\sigma^2} \Rightarrow \frac{\partial \ell}{\partial \vartheta} = \frac{x-\vartheta}{\sigma^2} = 0 \Rightarrow \hat{\vartheta} = x$$

$$\frac{\partial^2 \ell}{\partial \vartheta^2} = -\frac{1}{\sigma^2} \Rightarrow \sigma^2 = \frac{1}{I_F(\hat{\vartheta})}$$

More in general, the Fisher Information can be defined for any estimator T as:

$$I^T(\vartheta) \stackrel{d}{=} E\left(-\frac{\partial^2}{\partial \vartheta^2} \log f_T(X|\vartheta)\right)$$

↑ pdf chosen by the estimator T

It can be shown that: $I^T(\vartheta) \leq I_n(\vartheta)$, i.e., a generic estimator T will not bear more information than the random sample ⇒ MLE maximizes the information carried by the estimator (i.e., it reaches the upper bound).

* ML Estimators are efficient

T_n - efficient $\overset{DEF}{\Leftrightarrow}$ T_n - consistent and asymptotically normal AND

$$\frac{I_n(T_n)}{I_n(\vartheta^*)} \to 1 \text{ as } n \to \infty \Rightarrow \sqrt{I_n(\vartheta^*)}\,(T_n - \vartheta^*) \xrightarrow{D} N(0,1)$$

"Efficiency" means that the estimator contains the maximal amount of information supplied by the random sample about the value of the parameter ϑ^*

From a practical standpoint, we exploit the fact that $T_n \xrightarrow{P} \vartheta^* \Rightarrow \frac{I_n(T_n)}{I_n(\vartheta^*)} \to 1$

and consider the form: $\sqrt{I_n(T_n)}\,(T_n - \vartheta^*) \xrightarrow{D} (0,1)$

In this case, an approximation of the Fisher Information $I_n(T_n)$ is given

by the sampled (i.e., observed) numerical derivative $-\frac{\partial^2 \ell}{\partial \vartheta^2}(\bar{I}_n)$, i.e., we replace $I_n(\hat{\vartheta})$ with:

$$I_{obs}(\hat{\vartheta}) \triangleq -\frac{\partial^2 \ell}{\partial \vartheta^2}(\hat{\vartheta})$$

It can be shown that, as $n \to \infty$, the standard error of the MLE is:

$$SE = \frac{1}{\sqrt{I_{obs}(\hat{\vartheta})}}$$

and an approximate 95% confidence interval is $(\hat{\vartheta} - 2SE, \hat{\vartheta} + 2SE)$.

Ex.: $X_i \sim Exp(\lambda)$ $i = 1, 2, \dots, n$

$\vartheta \triangleq \lambda$ $\ell(\vartheta) = \log\left(\prod_{i=1}^{n} \vartheta e^{-\vartheta x_i}\right) = n \log \vartheta - n\vartheta \bar{x}$

with $\bar{x} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i$

We have: $\frac{\partial \ell}{\partial \vartheta} = \frac{n}{\vartheta} - n\bar{x} = 0 \iff \hat{\vartheta} = \frac{1}{\bar{x}}$

$$\frac{\partial^2 \ell}{\partial \vartheta^2} = -\frac{n}{\vartheta^2} \Rightarrow I_n(\hat{\vartheta}) = n\bar{x}^2 \Rightarrow SE = \frac{1}{\bar{x}\sqrt{n}}$$

Ex.: $X \sim B(n, p)$

$\vartheta \triangleq p$ $\ell(\vartheta) \triangleq \log\left(\vartheta^x (1-\vartheta)^{n-x}\right) = x \log \vartheta + (n-x)\log(1-\vartheta)$

$$\frac{\partial \ell}{\partial \vartheta} = \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta} = 0 \iff \hat{\vartheta} = \frac{x}{n}$$

$$\frac{\partial^2 \ell}{\partial \vartheta^2} = -\frac{x}{\vartheta^2} - \frac{n-x}{(1-\vartheta)^2} \Rightarrow I_n(\hat{\vartheta}) = \frac{n^2}{x} + \frac{n^2(n-x)}{(n-x)^2} = \frac{n^3}{x(n-x)}$$

$$= \frac{n}{\hat{\vartheta}(1-\hat{\vartheta})}$$

$$\Rightarrow SE = \sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}$$

* ML Estimators are approximately Bayesian

Let us consider the general form of the Bayes' theorem:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{\int_{-\infty}^{+\infty} f_{X|Y}(x|y) f_Y(y) dy}$$

Let us assume: $Y = T_n = \theta \Rightarrow f_{\theta|X}(\theta|x) = \frac{f_X(x|\theta) f_\theta(\theta)}{\int_{-\infty}^{+\infty} f_X(x|\theta) f_\theta(\theta) d\theta}$

Let us call: $\pi(\theta) \triangleq f_\theta(\theta) \triangleq$ "prior distribution" $\Rightarrow$ It represents our knowledge before seeing the data

$f_{\theta|X}(\theta|x) \triangleq$ "posterior distribution"

Because the likelihood function $L(\theta) \propto f_X(x|\theta)$, we can write:

$$f_{\theta|X}(\theta|x) = \frac{L(\theta)\pi(\theta)}{\int_{-\infty}^{+\infty} L(\theta)\pi(\theta) d\theta}$$

It can be proved that, for large samples, $f_{\theta|X}(\theta|x) \sim N\left(\hat{\theta}, \frac{1}{-\ell''(\hat{\theta})}\right)$

where $\ell''(\theta) \triangleq \frac{\partial^2 \ell}{\partial \theta^2}$

Finally, one can observe the following property:

$X_1 \sim N(\mu, \sigma_1^2) \quad X_2 \sim N(\mu, \sigma_2^2)$

In this case, we define:

$$\vartheta \triangleq \mu \qquad \ell(\vartheta) \triangleq -\frac{(x_1-\vartheta)^2}{2\sigma_1^2} - \frac{(x_2-\vartheta)^2}{2\sigma_2^2}$$

The MLE is obtained by imposing:

$$\frac{\partial \ell}{\partial \vartheta} = \frac{x_1-\vartheta}{\sigma_1^2} + \frac{x_2-\vartheta}{\sigma_2^2} = 0 \iff \hat{\vartheta} = \frac{\frac{1}{\sigma_1^2}x_1 + \frac{1}{\sigma_2^2}x_2}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}}$$

$\Rightarrow$ Denoted with $w_1 \triangleq \frac{1/\sigma_1^2}{1/\sigma_1^2+1/\sigma_2^2}$ $\quad w_2 \triangleq \frac{1/\sigma_2^2}{1/\sigma_1^2+1/\sigma_2^2}$, we have:

$T_N = w_1 X_1 + w_2 X_2 \Rightarrow T_N$ is the weighted mean of the RVs X_1 and X_2 □

References:

Textbook: ch.5 (section 5.1 and 5.2)

ch.7 (section 7.2 and 7.3)

ch.8 (section 8.1, 8.2, 8.3, and 8.4)