if I want to name one

MLE $L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$

$X_1, X_2 \dots X_n \overset{iid}{\sim} f(\cdot;\theta)$

if distribution is $\begin{cases} \text{continuous} & f(x;\theta) \\ \text{discrete} & p(x;\theta) / P(X=x;\theta) \end{cases}$

$X_1, X_2 \dots X_n \overset{iid}{\sim} N(\theta, 1)$ $f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$

$X_1, X_2 \dots X_n \overset{iid}{\sim} pois(\theta)$ $p(X=x;\theta) = \frac{\theta^x \cdot e^{-\theta}}{x!}$

Joint density $f(x_1, x_2 \dots x_n;\theta) \overset{iid}{=} f(x_1;\theta) f(x_2;\theta) \dots f(x_n;\theta)$

the "chance" of observing the data given one value of θ.

Likelihood of θ: defined as:

$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ (iid)

$\to$ n arguments, $x_1, x_2 \dots x_n$ (given fixed θ)

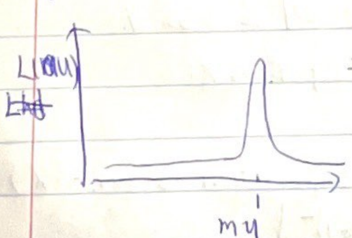
one argument

treat $x_1, x_2 \dots x_n$ as fixed.

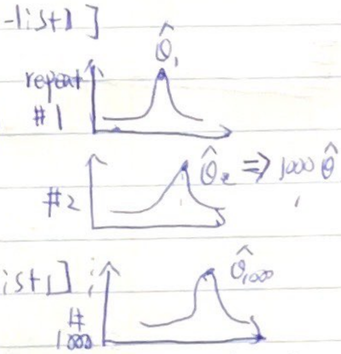
MLE of θ: $\hat{\theta}_{MLE}$ = Maximizer of $L(\theta)$ = $\underset{\theta}{\arg\max}\, L(\theta)$

: argument value that max fun $L(\theta)$

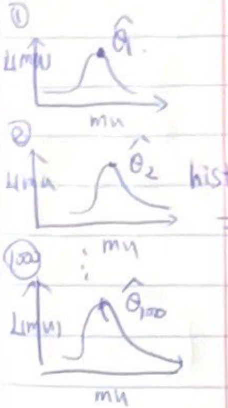
```
mu=1, n=100   X<-rnorm(n, mean=mu)
mu-list <- seq(-2, 2, 0.01)  [my guesses of mu]
lhd_list <- sapply(mu-list, FUN=function(mu-cand){
    prod(dnorm(x, mean=mu-cand))  # likelihood under diff mu/over mu-list
})  # likelihood is the product of n densities/ "prob of X around xi"
plot(x=mu-list, y=lhd-list) ; mu-list[which.max(lhd-list)]
```

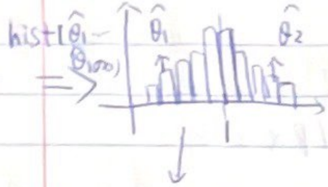



```
# repeat 1000 times.
=> sapply(1:1000, FUN=function(i){
     X<- rnorm(n, mean=mu)
     :
     mu_list[which.max(lhd-list)];
})
```



[each time, $x_1, x_2 \dots x_n$ are changing, so $L(\theta)$ changing]





$\hat{\theta}_{MLE}$ will change if data $(x_1, x_2 \dots x_n)$ changes
So θ is a RV as $X_1 \dots X_n$ are RVs ☆

distribution of $\hat{\theta}$.

[Likelihood function changes with RV. X_s]

For one sample, $\hat{\theta}$ may be away from the true θ, but stats is talking abt overall distribution (not one particular trial). So, on average $\hat{\theta}$ is close to true θ which is 1 in this example.

intuitively, MLE is a Normal distribution

Properties of $\hat{\theta}$

① bias : $bias(\hat{\theta}) = E(\hat{\theta}) - \theta$ [$\hat{\theta}$ is a RV. we look at the expectation]
② variance $var(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$ by def
③ MSE (Mean squared error) $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

Cramer-Rao Inequality (Theorem) : the minimum variance of unbiased $\hat{\theta}$.

Let $X_1, X_2 \dots X_n \overset{iid}{\sim} \text{pdf } f(\cdot;\theta)$;
Let $\hat{\theta} = g(X_1, X_2 \dots X_n)$ be an unbiased estimator of θ.
Then under smoothness conditions of $f(\cdot;\theta)$ [differentiable]

$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$ where $I(\theta)$ is Fisher information

Fisher information : based on the log-likelihood of θ w/ 1 obs. X_i

$L(\theta) = \log f(X_i;\theta)$ f is the pdf/pmf of r.v. X_i

$$I(\theta) = E[(\frac{d}{d\theta} L(\theta))^2]$$ (θ is a fixed, expectation is w.r.t. X)

$$= E_X[(\frac{d}{d\theta} L(\theta))^2] = E_X[(\frac{d}{d\theta} \log f(x,\theta))^2]$$

evaluate how much information does the density provide abt θ.
It's a property of the distribution alone not depending on real data (Expectation)

一阶导数：函数值变化大.

二阶导数=(函数的)斜率变化大, 曲率↑

Alternative def: $I(\theta) = -\underset{X}{E}\left[\frac{d^2}{d\theta^2} L(\theta)\right]$

so when θ is given, $L(\theta)$ is known. $L(\theta)$ may or may not depend on θ.

proof. $\int f(x;\theta)\,dx = 1$ [def $L(\theta) = \log f(X;\theta)$]

$$\frac{d}{d\theta}\int f(x;\theta)\,dx = \frac{d}{d\theta} 1 = 0.$$

because of the smoothness condition of $f(\cdot;\theta)$

swap $\frac{d}{d\theta}$ and $\int$: $\int \frac{d}{d\theta} f(x;\theta)\,dx = \frac{d}{d\theta} 1 = 0.$ ①

$$\frac{d}{d\theta}\log f(x;\theta) = \frac{\frac{d}{d\theta} f(x;\theta)}{f(x;\theta)} \quad \text{[chain rule]}$$

$$\therefore \frac{d}{d\theta} f(x;\theta) = \left(\frac{d}{d\theta}\log f(x;\theta)\right) \cdot f(x;\theta)$$

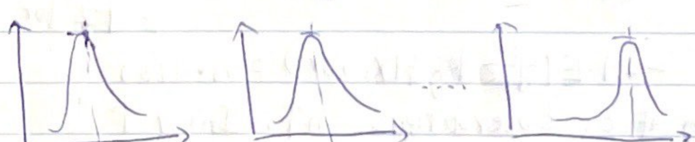
apply $\int \cdot\, dx$ to both sides:

$$\underbrace{\int \frac{d}{d\theta} f(x;\theta)\,dx}_{① \;\to\; 0} = \int \underbrace{\left(\frac{d}{d\theta}\log f(\overset{\text{only value a RV takes}}{x};\theta)\right)}_{\text{a fun of } X} f(x;\theta)\,dx$$

$$0 = E\left(\frac{d}{d\theta}\log f(\underset{\text{RV}}{X};\theta)\right) \quad \text{[by def of expectation]}$$

Intuitively, FI indicates "how fast is the likelihood changing at current value of θ." [curvature]

FI is the average of first derivative over all Xs.



#① $X^{\#1}$ #②: $X^{\#2}$ ... #n $X^{\#n}$

$I(\theta)\uparrow \rightarrow \frac{dL(\theta)}{d\theta}\uparrow \rightarrow$ change θ a little bit, $L(\theta)$ changes alot / the most.

$\rightarrow$ θ_{true} is a sensitive spot ~~to~~ of Likelihood function $\rightarrow$ diff~~erentiate~~ $\hat{\theta}$ from other θs. $I(\theta_{true})$ is the max.

$$\frac{d}{d\theta} f(x;\theta) = \left(\frac{d}{d\theta}\log f(x;\theta)\right)\cdot f(x;\theta) \quad ②$$

$(ab)' = a'b + b'a$

take derivative on ② $\frac{d^2}{d\theta^2} f(x;\theta) = \left(\frac{d^2}{d\theta^2}\log f(x;\theta)\right)\cdot f(x;\theta) + \frac{d}{d\theta}\log f(x;\theta)\left(\cdot\underbrace{\frac{d}{d\theta} f(x;\theta)}_{②}\right)$

apply $\int \cdot\, dx$ integration $\int \frac{d^2}{d\theta^2} f(x;\theta)\,dx \underset{\text{smothness}}{=} \frac{d^2}{d\theta^2}\underbrace{\int f(x,\theta)\,dx}_{1} = \frac{d^2}{d\theta^2}1 = 0$ [LHS]

[RHS]: $\underbrace{\int \frac{d^2}{d\theta^2}\log f(x;\theta)\cdot f(x;\theta)\,dx}$ $+$ $\underbrace{\int \frac{d}{d\theta}\log f(x;\theta)\cdot\left(\frac{d}{d\theta}\log f(x,\theta)\right)\cdot f(x;\theta)\,dx}$

$$= E\left(\frac{d^2}{d\theta^2}\log f(X;\theta)\right] + E\left(\left(\frac{d}{d\theta}\log f(X;\theta)\right)^2\right]$$

$$LHS = 0 = RHS = E_X\left(\frac{d^2}{d\theta^2}\log f(X;\theta)\right) + E_X\left(\left(\frac{d}{d\theta}\log f(X;\theta)\right)^2\right]$$

<u>Joint Fisher Information</u> $X_1, X_2 \dots X_n \overset{iid}{\sim} f(\cdot;\theta)$

$$I_n(\theta) = E\left[\left[\frac{d}{d\theta}\log f(X_1, X_2, \dots X_n;\theta)\right]^2\right]$$

$$= E\left[\left[\frac{d}{d\theta}\sum_{i=1}^{n}\log f(X_i;\theta)\right]^2\right] \quad [\text{iid}]$$

$$I_n(\theta) = -E\left[\frac{d^2}{d\theta^2}\log f(X_1, X_2 \dots X_n;\theta)\right]$$

$$= -E\left(\frac{d^2}{d\theta^2}\sum_{i=1}^{n}\log f(X_i;\theta)\right)$$ [put derivative inside sum]

$$= -E\left(\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f(X_i;\theta)\right)$$

$$= -\sum_{i=1}^{n} E\left(\frac{d^2}{d\theta^2}\log f(X_i;\theta)\right)$$

FI for one RV.

$$= -n\cdot E\left(\frac{d^2}{d\theta^2}\log f(X_i;\theta)\right) = n\cdot I(\theta)$$

"FI scaled with # of observations, $n\uparrow$, $I_n(\theta)\uparrow$."

<u>Cramer-Rao low bound</u>: $\hat{\theta} = g(X_1, X_2 \dots X_n)$ is unbiased, $X_1, X_2 \dots X_n \overset{iid}{\sim} f(\cdot,\theta)$

Then $Var(\hat{\theta}) \geq \frac{1}{nI(\theta)} = \frac{1}{I_n(\theta)}$

correlation $cor(X,Y) = \dfrac{cov(X,Y)}{\sqrt{varX}\sqrt{varY}}$ $[-1, 1]$

$$cor^2(X,Y) = \frac{cov^2(X,Y)}{var(X)var(Y)} \quad [0,1] \quad \le 1.$$

Cauchy-Schwartz $cov^2(X,Y) \le var(X)\,var(Y)$

proof of Cramer-Rao: use some RV T, which satisfy:

$$Var(\hat{\theta}) \geq \frac{cov^2(\hat{\theta},T)}{Var(T)} \overset{\text{for some } T}{=\!=} \frac{1}{nI(\theta)}.$$

based on previous knowledge: Let $T = \frac{d}{d\theta}\log f(X_1, X_2 \ldots X_n;\theta) = \frac{d}{d\theta}\sum_{i=1}^{n}\log f(X_i;\theta)$

$$= \sum_{i=1}^{n}\frac{d}{d\theta}\log f(X_i;\theta)$$

$$= \sum_{i=1}^{n}\frac{\frac{d}{d\theta}f(X_i;\theta)}{f(X_i;\theta)}$$

$$var(T) = E[T^2] - (ET)^2.$$

$$= \underbrace{E\left[\left(\frac{d}{d\theta}\log f(X_1 \ldots X_n;\theta)\right)^2\right]}_{\text{def of } I_n(\theta)} - (ET)^2 = I_n(\theta) - (ET)^2$$

$$= I_n(\theta) = nI(\theta).$$

$$E(T) = \int\cdots\int \sum_{i=1}^{n}\frac{\frac{d}{d\theta}f(x_i;\theta)}{f(x_i;\theta)} f(x_1, x_2 \ldots x_n)\,dx_1 \ldots dx_n$$

$$= \int\cdots\int\left(\sum_{i=1}^{n}\frac{\frac{d}{d\theta}f(x_i;\theta)}{f(x_i;\theta)}\right) f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta)\,dx_1dx_2\ldots dx_n.$$

$$= 0 \;/\; E(T) = E\left(\sum_{i=1}^{n}\frac{\frac{d}{d\theta}f(x_i;\theta)}{f(x_i;\theta)}\right) \overset{iid}{=} nE\left(\frac{\frac{d}{d\theta}f(x_i;\theta)}{f(x_i;\theta)}\right) \overset{\text{shown before}}{=} n\cdot 0 = 0$$

$$cov(\hat{\theta},T) = E(\hat{\theta}T) - E(\hat{\theta})\underbrace{E(T)}_{0} = E(\hat{\theta}T)$$

$$E(\hat{\theta}T) \overset{\text{be def}}{=} \int\cdots\int g(x_1, x_2 \ldots x_n)\left(\sum_{i=1}^{n}\frac{\frac{d}{d\theta}f(x_i;\theta)}{f(x_i;\theta)}\right)\left(\prod_{i=1}^{n} f(x_i;\theta)\right)dx_1 \ldots dx_n$$