Ch7.5 Maximum likelihood estimators (MLEs) θ fixed, not random

Maximum likelihood estimation relies on likelihood functions while avoiding the use of prior distributions and loss functions.

Given a likelihood function, it determines the value of θ that maximizes it.

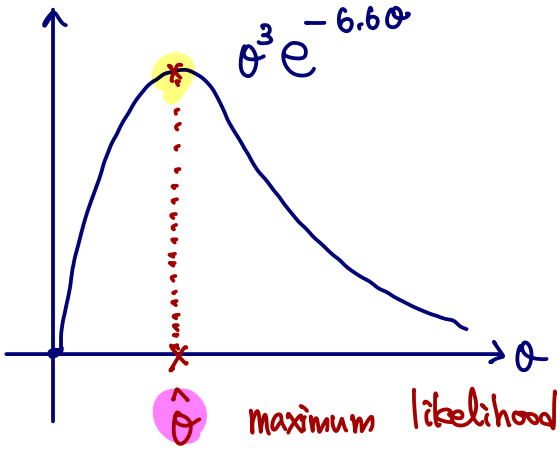
Let $X_1, \dots, X_n$ be a random sample from a distribution with $f(x|\theta)$ where θ is unknown. For fixed $\boldsymbol{x} = (x_1, \dots, x_n)$ (observed data), the joint probability distribution $f_n(\boldsymbol{x}|\theta)$ is called the likelihood function when it is regarded as a function of θ given the observed data $x_1, \dots, x_n$.

It quantifies how likely different values of θ, given the data.

* Aim to find a value of θ that is most consistent with data.

EG. [Lifetimes of Electronic Components] Lifetimes $X_1, \dots, X_n$ of electronic components are modeled as iid exponential random variables with parameter θ. The observed data are $x_1 = 3, x_2 = 1.5, x_3 = 2.1$.

$$f_3(x_1, x_2, x_3 | \theta) \overset{\text{indep}}{=} \prod_{i=1}^{3} f(x_i|\theta) \overset{\text{iid}}{=} \prod_{i=1}^{3} \theta e^{-\theta x_i} = \theta^3 e^{-\theta \sum_{i=1}^{3} x_i} = \theta^3 e^{-6.6\theta}$$



$\hat{\theta}$ maximum likelihood estimate.

With $\hat{\theta}$, can explain data best in terms of likelihood under the statistical model "iid exponential".

Definition. For each $\boldsymbol{x}$, let $\delta(\boldsymbol{x}) \in \Omega$ be such that $\delta(\boldsymbol{x})$ maximizes the likelihood function $f_n(\boldsymbol{x}|\theta)$.

← MLE.

$\hat{\theta} = \delta(\boldsymbol{x})$ is called the maximum likelihood estimate of θ.

$\hat{\theta} = \delta(\boldsymbol{X})$ is called the maximum likelihood estimator of θ.

g^* ← Bayes estimators.

To simplify computations (especially when dealing with products of probabilities), the logarithm of the likelihood function is often used.

Definition. Call $L(\theta) = \log f_n(\boldsymbol{x}|\theta)$ the log likelihood function. (natural log.)

Fact: The log function is monotonically increasing on $(0, \infty)$

value of θ that maximizes $L(\theta)$

$\Vert$

value of θ that maximizes $f_n(\boldsymbol{x}|\theta)$

Find θ that maximizes $L(\theta) = \hat{\theta}$ MLE.

EG. [Lifetimes of Electronic Components]

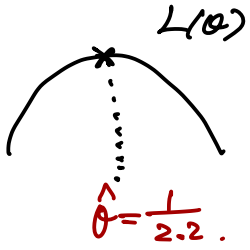
$$f_3(x_1, x_2, x_3|\theta) = \theta^3 e^{-6.6\theta}.$$

$$\Rightarrow L(\theta) = \log f_3(x_1, x_2, x_3|\theta) = \log \theta^3 e^{-6.6\theta}.$$

$$= \log \theta^3 + \log e^{-6.6\theta} = 3\log\theta - 6.6\theta.$$

Find θ that maximizes $L(\theta) = 3\log\theta - 6.6\theta$!!

$$\Rightarrow \frac{\partial L(\theta)}{\partial \theta} = \frac{3}{\theta} - 6.6 \overset{\text{set}}{=} 0 \quad \Rightarrow \theta = \frac{3}{6.6} = \frac{1}{2.2}$$



Since $\left.\frac{\partial^2 L(\theta)}{\partial \theta^2}\right|_{\theta=\frac{1}{2.2}} = \left.-\frac{3}{\theta^2}\right|_{\theta=\frac{1}{2.2}} = -\frac{3}{\frac{1}{2.2^2}} < 0$,

$$\hat{\theta} = \frac{1}{2.2}.$$

EG. Given $X_1, \dots, X_n|\theta \sim$ iid Bernoulli θ, where $0 \le \theta \le 1$, find the MLE $\hat{\theta}$.

$$f_n(x_1, \cdots, x_n|\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}$$

$$\Rightarrow L(\theta) = \log \theta^{\sum x_i} \cdot (1-\theta)^{n-\sum x_i}$$

$$= \log \theta^{\sum x_i} + \log (1-\theta)^{n-\sum x_i}$$

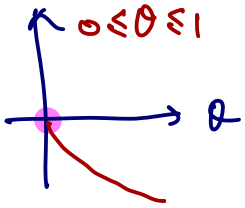
$$= \sum x_i \cdot \log\theta + (n-\sum x_i) \log(1-\theta).$$

1) $\sum x_i \notin \{0, n\}$: $\frac{\partial L(\theta)}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n-\sum x_i}{1-\theta} \overset{\text{set}}{=} 0$ $\times \theta(1-\theta)$

$$\Rightarrow (1-\theta)\sum x_i - \theta(n-\sum x_i) = 0$$

$$\Rightarrow \sum x_i - \sum x_i \theta - n\theta + \sum x_i \theta = 0$$

$$\Rightarrow \sum x_i = n\theta \Rightarrow \theta = \frac{\sum x_i}{n} = \bar{x}_n.$$

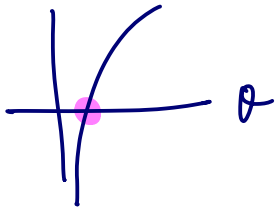


2) $\sum x_i = 0$; $L(\theta) = n \log(1-\theta)$ is decreasing

$$\hat{\theta} = 0 \ (= \bar{x}_n)$$

3) $\sum x_i = n$; $L(\theta) = n \log\theta$ is increasing

$$\hat{\theta} = 1 \ (= \bar{x}_n)$$



$$\Rightarrow \hat{\theta} = \bar{X}_n.$$

$X_1, X_2, \cdots, X_n \mid \theta \sim$ iid w/ $f(x \mid \theta)$ ← unknown, fixed.

Aim to construct an estimator for θ.
a function of $X_1, \cdots, X_n$

In MLE, $\hat{\theta}$ that maximizes $f_n(x_1, \cdots, x_n \mid \theta)$
likelihood function.
a function of θ given $x_1, \cdots, x_n$.

$\hat{\theta}$ is most consistent w/ data under the likelihood function considered.

EG. Given $X_1, \dots, X_n | \mu, \sigma^2 \sim$ iid normal $\theta = (\mu, \sigma^2)$, where $-\infty < \mu < \infty$, $\sigma^2 > 0$, find the MLE $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$

↪ unknown.

$$f_n(x_1, \dots, x_n | \mu, \sigma^2) \overset{\text{iid}}{=} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

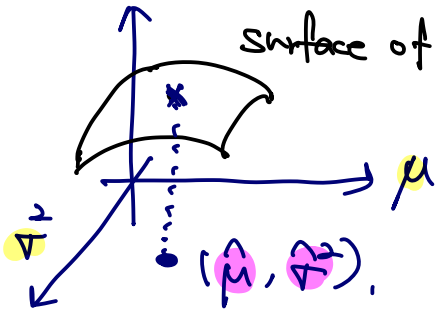
$$= \frac{1}{\sqrt{2\pi}^n} \times \frac{1}{\sqrt{\sigma^2}^n} \times e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$= (\sigma^2)^{-\frac{n}{2}}$

$$\Rightarrow L(\mu, \sigma^2) = \log \frac{1}{\sqrt{2\pi}^n} + \log \frac{1}{\sqrt{\sigma^2}^n} + \log e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= c - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Note that L is smooth and has the maximum point (verified using 2nd order derivatives and Jacobians)



① $\frac{\partial L(\mu, \sigma^2)}{\partial \mu} = 0$, ② $\frac{\partial L(\mu, \sigma^2)}{\partial \sigma^2} = 0$.

$\times\sigma^2$ $\times\sigma^2$

① $\frac{\partial L(\mu, \sigma^2)}{\partial \mu} = 0 - 0 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu) \overset{\text{set}}{=} 0 \Rightarrow \sum(x_i-\mu) = 0$

$\Rightarrow \sum x_i - n\mu = 0 \Rightarrow \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}_n$

$\times 2\sigma^4$ $\times 2\sigma^4$

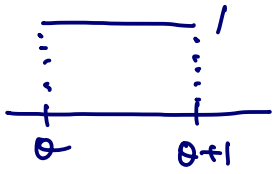
② $\frac{\partial L(\mu, \sigma^2)}{\partial \sigma^2} = 0 - \frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(x_i-\mu)^2 \overset{\text{set}}{=} 0$

$-n\sigma^2 + \sum(x_i-\mu)^2 = 0 \Rightarrow \sigma^2 = \frac{1}{n}\sum(x_i-\mu)^2$.

$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum(x_i-\bar{x}_n)^2 \neq$ sample variance $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$

Note MLEs aren't always unique. MLEs don't always exist ;(

Page 423 $X_1, \ldots, X_n|\theta \sim$ iid continuous uniform on $[\theta, \theta+1]$



$$f(x|\theta) = 1 \times I(\theta \le x \le \theta+1)$$

$$\Rightarrow f_n(x_1,\cdots,x_n|\theta) = \prod_{i=1}^{n} 1 \times I(\theta \le x_i \le \theta+1)$$

$$= I(\theta \le x_1,\cdots,x_n \le \theta+1)$$

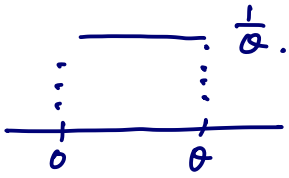
$$= I(\theta \le \min\{x_1,\cdots,x_n\} \le \max\{x_1,\cdots,x_n\} \le \theta+1)$$

$$\max\{x_1,\cdots,x_n\}-1 \le \theta \le \min\{x_1,\cdots,x_n\}$$

The likelihood function can be maximized at any values btw $\max\{x_1,\cdots,x_n\}-1$ and $\min\{x_1,\cdots,x_n\}$

$\Rightarrow$ MLEs are not unique.

Page 422 $X_1, \ldots, X_n|\theta \sim$ iid continuous uniform on $(0, \theta)$



$$f(x|\theta) = \frac{1}{\theta} \times I(0 < x < \theta)$$

$$\Rightarrow f_n(x_1,\cdots,x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 < x_i < \theta)$$

$$= \frac{1}{\theta^n} I(0 < x_1,\cdots,x_n < \theta)$$

$$= \frac{1}{\theta^n} I(0 < \min x_i \le \max x_i < \theta) \quad (*)$$

Pick θ as small as possible subject to $(*)$

There exists no such θ because

θ cannot be chosen equal to $\max\{x_1,\cdots,x_n\}$.

$\Rightarrow$ MLEs of θ do not exist!

The MLEs for many of the Ch5 special distributions

unknown parameter.

$$X_1, \ldots, X_n | \theta \sim \text{iid } f(x|\theta)$$

distribution, $f(x|\theta)$	MLE $\hat{\theta}$
Bernoulli θ ← prob of success	$\hat{\theta} = \bar{X}_n$ ← proportion of successes. (lecture note).
Poisson θ ← average # of events occurring	$\hat{\theta} = \bar{X}_n$ #9 in Ch 3.3
exponential θ ← failure rate = 1 / average.	$\hat{\theta} = \frac{1}{\bar{X}_n}$ #3
uniform $[a, \theta]$ ← end point.	$\hat{\theta} = \max\{X_1, \ldots X_n\}$. } #11.
uniform $[\theta, b]$ ← starting point	$\hat{\theta} = \min\{X_1, \ldots X_n\}$ }
normal μ, σ^2, both unknown	$\hat{\mu} = \bar{X}_n$
	$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ } lecture note.
normal θ, σ^2, mean unknown	$\hat{\theta} = \bar{X}_n$
normal μ, θ, variance unknown	$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$

Ch7.6 Properties of MLEs

1. invariance property

failure rate = $\frac{1}{\text{average}}$

Q. $X_1, \dots, X_n|\theta \sim$iid exponential θ. What is the MLE of θ^2?

$\hat{\theta}$ = MLE of θ = $\frac{1}{\bar{X}_n}$.

Can we say $\left(\frac{1}{\bar{X}_n}\right)^2$ is an MLE of θ^2? Yes by invariance thm.

Theorem (Invariance). Let $\hat{\theta}$ be an MLE of θ, and let $g(\theta)$ be a function of θ, then an MLE of $g(\theta)$ is $g(\hat{\theta})$.

$P(X=x) = \frac{e^{-\theta}\cdot\theta^x}{x!}$ ← $x=0,1,2,\dots$

EG. $X_1, \dots, X_n|\theta \sim$ iid Poisson θ. What is the MLE of $P(X_1 = 0|\theta) = e^{-\theta}$?

$\hat{\theta} = \bar{X}_n$

MLE of $e^{-\theta} = g(\theta) \Rightarrow e^{-\hat{\theta}} = e^{-\bar{X}_n}$.

EG. $X_1, \dots, X_n|\ \mu, \sigma^2 \sim$ iid normal μ, σ^2. What the an MLE of σ and $E[X_1^2]$?

$\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$\sigma^2 = \mathrm{Var}[X_1] + (E[X_1])^2$

$\sigma^2 + \mu^2$

* $\sigma = \sqrt{\sigma^2} \Rightarrow \hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$

* $\sigma^2 + \mu^2 \Rightarrow \hat{\sigma}^2 + (\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + (\bar{X}_n)^2$.

2. Asymptotic behaviors ($n \to \infty$) behavior of MLEs when n is large.

Fact: Under some regularity conditions (often hold in practice), MLEs are

1. consistent for true θ_0, i.e., $\hat{\theta} \xrightarrow{P} \theta_0$, as $n \to \infty$.

2. asymptotically normal, i.e., $\hat{\theta} \approx$ normal w/ θ_0, $\frac{1}{n I(\theta_0)}$,

   where $I(\theta_0)$ is Fisher Information (in Ch8).

Method of moments (MoM)

It is a simple technique for estimating parameters. It match empirical (sample) moments (data.) to the population moments (theoretical values) of a distribution.

Assume that $X_1, \ldots, X_n$ form a random sample from a distribution indexed by a k-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. ex) normal μ, σ^2 2-dimensional ($\mu = \theta_1$, $\sigma^2 = \theta_2$)

Definition.

Define the j^{th} sample moments by $m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$, for $j = 1,2, \ldots$.

Define the j^{th} population moments by $\mu_j = E[X_1^j]$, for $j = 1,2, \ldots$.

The method of moments estimator, denoted by $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ for θ, is the solution to the system of k equations (k unknown parameters).

$$\begin{cases} m_1 = \mu_1 \\ m_2 = \mu_2 \\ \vdots \\ m_k = \mu_k \end{cases}$$

Solve for $\theta = (\theta_1, \ldots, \theta_k)$

EG $X_1, \ldots, X_n | \mu, \sigma^2 \sim$ iid normal μ, σ^2 (θ_1 θ_2; two unknown parameters.)

Sample moment

$$m_1 = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$$

$$m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$$

population moment

$$\mu_1 = E[X_1] = \mu$$

$$\mu_2 = E[X_1^2] = Var[X_1] + (E[X_1])^2 = \sigma^2 + \mu^2$$

$$\Rightarrow \begin{cases} m_1 = \mu_1 \\ m_2 = \mu_2 \end{cases} \Rightarrow \begin{cases} \bar{X}_n = \mu \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = \sigma^2 + \mu^2 \end{cases} \Rightarrow \text{Solve for } \mu \text{ and } \sigma^2 !$$

$$\Rightarrow \tilde{\mu} = \bar{X}_n \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 \overset{(*)}{=} \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

($\tilde{\mu}$: MoM for μ; $\tilde{\sigma}^2$: MoM for σ^2)

$= \hat{\mu}$ (MLE for μ) $\qquad = \hat{\sigma}^2$ (MLE for σ^2)

$$(*)\ \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n (X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2) = \frac{1}{n}\left[\sum X_i^2 - 2\bar{X}_n \sum X_i + n\bar{X}_n^2\right]$$

$$= \frac{1}{n}\sum X_i^2 - 2(\bar{X}_n)^2 + \bar{X}_n^2 = \frac{1}{n}\sum X_i^2 - (\bar{X}_n)^2.$$

Note.

- In some cases, MLE = MoM (see the previous example)
- When MLE ≠ MoM, generally MLE is considered "better"

because MLEs are asymptotically ~~efficient~~ ! ($n \to \infty$) (small variance)

⇒ Roughly speaking, MLEs are more precise when n is large.

- But MoM is typically easy to compute. ← when you need a quick estimate!

EG $X_1, \ldots, X_n | \alpha \sim$ iid gamma α, 1 $\longrightarrow f(x|\alpha) = \frac{1^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \; x > 0$.

MoM: $m_1 = \bar{x}_n$ and $\mu_1 = E[X_1] = \frac{\alpha}{1} = \alpha$.

$\Rightarrow m_1 = \mu_1 \Rightarrow \tilde{\alpha} = \bar{x}_n$ ← easy to derive!

MLE: $f_n(x_1, \cdots, x_n | \alpha) = \prod_{i=1}^n \frac{1}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i} = \frac{1}{\Gamma(\alpha)^n} \left(\prod x_i\right)^{\alpha-1} e^{-\sum x_i}$

$\Rightarrow L(\alpha) = \log \frac{1}{\Gamma(\alpha)^n} + \log \left(\prod x_i\right)^{\alpha-1} + \log e^{-\sum x_i}$

$= -n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^n \log x_i - \sum_{i=1}^n x_i$.

$\Rightarrow \frac{dL(\alpha)}{d\alpha} = -n \times \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log x_i \overset{\text{set}}{=} 0$.

$\Rightarrow \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{1}{n} \sum_{i=1}^n \log x_i = 0$.

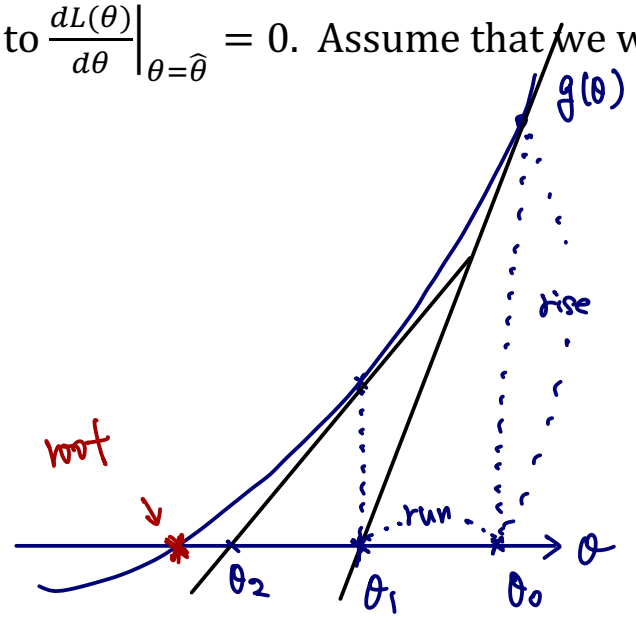
Need to solve for α to get MLE $\hat{\alpha}$!

no closed form solution!

⇒ Numerical method is needed!

Newton's Method

Numerical analysis is often needed to compute MLEs. When there is only one parameter, **Newton's method** offers a way to numerically compute a root to $\left.\frac{dL(\theta)}{d\theta}\right|_{\theta=\widehat{\theta}} = 0$. Assume that we want to find the root of $g(\theta) = 0$.



1. Pick an initial guess θ_0
2. Iterate $\theta_{k+1} = \theta_k - \frac{g(\theta_k)}{g'(\theta_k)}$, $k = 0, 1, 2, \cdots$

The slope at $\theta_0 = g'(\theta_0) = \frac{\text{rise}}{\text{run}} = \frac{g(\theta_0)}{\theta_0 - \theta_1}$

$\Rightarrow \theta_0 - \theta_1 = \frac{g(\theta_0)}{g'(\theta_0)} \Rightarrow \theta_1 = \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)}$

Back to EG Let X_1, X_2, X_3 be $\sim$ iid gamma $\alpha, 1$.

We wanted to solve $\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{1}{n}\sum_{i=1}^{n} \log x_i = 0$ for α.

$g(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{1}{n}\sum_{i=1}^{n} \log x_i = \text{digamma}(\alpha) - \frac{1}{n}\sum_{i=1}^{n} \log x_i$.

$\text{digamma}'(\alpha) = \text{trigamma}(\alpha)$.

1. $\alpha_0 = \bar{x}_n$ (MoM!)
2. $\alpha_{k+1} = \alpha_k - \frac{g(\alpha_k)}{g'(\alpha_k)} = \alpha_k - \frac{\text{digamma}(\alpha_k) - \frac{1}{n}\sum_{i=1}^{n} \log x_i}{\text{trigamma}(\alpha_k)}$,

$k = 0, 1, 2, 3 \cdots$

R output: If $(x_1, x_2, x_3) = (3,5,7)$, then MLE of α is $\hat{\alpha} = 5.208906$.

```
> x=c(3,7,5)
> mean(log(x))
[1] 1.55132
> (alpha=mean(x))
[1] 5
> (alpha=alpha-(digamma(alpha)-mean(log(x)))/trigamma(alpha))
[1] 5.204238
> (alpha=alpha-(digamma(alpha)-mean(log(x)))/trigamma(alpha))
[1] 5.208903
> (alpha=alpha-(digamma(alpha)-mean(log(x)))/trigamma(alpha))
[1] 5.208906
> (alpha=alpha-(digamma(alpha)-mean(log(x)))/trigamma(alpha))
[1] 5.208906
```

$= \frac{1}{n}\sum_{i=1}^{3} \log x_i = 1.55132$.

$\alpha_0 = \bar{x}_3$

α_1, α_2, α_3, α_4

$k=0$ ↙ recursive formula.

$k=1$, $k=2$, $k=3$.

EM algorithm

The **EM algorithm** is an iterative method for approximating MLEs when missing data (e.g., X_3 =?) or partial information (e.g., $X_2 < 8$) are making it difficult to find the MLE's closed form.

$f(x|\mu) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}, \; x > 0.$

EG. Let $X_1, X_2, X_3, X_4, \ldots$ be iid random variables from an exponential distribution with mean μ. Items are tested sequentially, and the study ends at time = 19. What if $x_1 = 3.1$, $x_2 = 6.9$, and $X_3 > 9$?

1. Initialize $\theta^{(0)}$ at a starting point (typically MoM is used).

$\mu^{(0)} = \bar{x}_2 = \frac{3.1 + 6.9}{2} = 5.$ $\quad x_1, x_2$

2. Write the *full-data-log-likelihood* $L(\theta)$. Plugin all non-missing values and represent all missing values as x_i. $\quad x_3$

$$L(\mu) = \log f_3(x_1, x_2, x_3|\mu) = \log \prod_{i=1}^{3} \frac{1}{\mu} e^{-\frac{1}{\mu}x_i} = \log \mu^{-3} e^{-\frac{1}{\mu}\sum_{i=1}^{3} x_i}$$

$$= -3\log\mu - \frac{1}{\mu}\sum_{i=1}^{3} x_i = -3\log\mu - \frac{1}{\mu}(3.1 + 6.9 + x_3).$$

3. **Expectation-Step:** Replace the missing value in $L(\theta)$ by its conditional mean given $\theta = \theta^{(k)}$ and all available inormation. Treat missing values as random.

$E[X_3 \mid X_3 > 9, \mu^{(0)} = 5]$

$= E[9 + X_3 - 9 \mid X_3 > 9, \mu^{(0)} = 5]$

Memoryless property of Exponential

$X_3 > 9 + t \mid X_3 > 9 \overset{d}{=} X_3 > t$

$\Rightarrow X_3 - 9 > t$

$= 9 + E[X_3 - 9 \mid X_3 > 9, \mu^{(0)} = 5] = 9 + E[X_3 \mid \mu^{(0)} = 5] = 9 + 5 = 14.$ $\quad = x_3^{(1)}$

4. **Maximization-Step:** The imputed likelihood $L(\theta)$ from Step 3 has no missing values and is just a function of θ. Maximize $L(\theta)$ to get $\theta^{(k+1)}$.

$$L(\mu) = -3\log\mu - \frac{1}{\mu}(3.1 + 6.9 + 14) = -3\log\mu - \frac{24}{\mu}.$$

$$\frac{dL(\mu)}{d\mu} = -\frac{3}{\mu} + \frac{24}{\mu^2} = 0 \Rightarrow \mu^{(1)} = 8$$

5. Stop if $\theta^{(k+1)} \approx \theta^{(k)}$ and go to Step 3 with $k = k + 1$ otherwise.

Step 3 $E[X_3 \mid X_3 > 9, \mu^{(1)} = 8] = 9 + 8 = 17 \leftarrow x_3^{(2)}$

Step 4 $\ell(\mu) = -3\log\mu - \frac{1}{\mu}(3.1 + 6.9 + 17) \Rightarrow \mu^{(2)} = 9$

R output:

```
> x=c(3.9,6.1,NA)
> (mu=(3.9+6.1)/2) #"mu zero"
[1] 5 = μ(0)
> for(i in 1:50) #START EM ALGORITHM
+ {
+     x[3]=9+mu
+     mu=mean(x)
+
+
+     if(i<4 || i>48){
+      print(paste("i=",i))
+      print(x)
+      print(mu)
+      print("")}
+ }
```

EM

```
[1] "i= 1"                 // x3(1)
[1]  3.9  6.1 14.0
[1] 8 = μ(1)
[1] ""
[1] "i= 2"                 // x3(2)
[1]  3.9  6.1 17.0
[1] 9 = μ(2)
[1] ""
[1] "i= 3"
[1]  3.9  6.1 18.0
[1] 9.333333
[1] ""
[1] "i= 49"                // x3(49)
[1]  3.9  6.1 18.5
[1] 9.5 = μ(49)
[1] ""
[1] "i= 50"                // x3(50)
[1]  3.9  6.1 18.5
[1] 9.5 = μ(50)
[1] ""
```

$\Rightarrow \hat{\mu} = 9.5$!!

Historical Facts: Dr. Jeff Wu (https://en.wikipedia.org/wiki/C.F._Jeff_Wu) visited WVU on April 25, 2017 to give a talk. Dr. Wu is known for his work on the convergence of the EM-algorithm in 1983, and resampling methods such as the bootstrap. He is also credited with popularizing the term "data science."

The Annals of Statistics
1983, Vol. 11, No. 1, 95–103

ON THE CONVERGENCE PROPERTIES OF THE EM ALGORITHM[1]

BY C. F. JEFF WU

University of Wisconsin, Madison

Two convergence aspects of the EM algorithm are studied: (i) does the EM algorithm find a local maximum or a stationary value of the (incomplete-data) likelihood function? (ii) does the sequence of parameter estimates generated by EM converge? Several convergence results are obtained under conditions that are applicable to many practical situations. Two useful special cases are: (a) if the unobserved complete-data specification can be described by a curved exponential family with compact parameter space, all the limit points of any EM sequence are stationary points of the likelihood function; (b) if the likelihood function is unimodal and a certain differentiability condition is satisfied, then any EM sequence converges to the unique maximum likelihood estimate. A list of key properties of the algorithm is included.

Ch7.7 Sufficient statistics

What if Bayes estimators, MLE, and MoM are unavailable?

R.A. Fisher introduces the concept of sufficient statistics, which provides a method for finding a good estimator for θ.

↙ function of $X_1, \cdots, X_n$.

Concept: We seek statistics $T(X_1, \dots, X_n)$ that are **sufficient** for finding a good estimator of θ in the sense that

1. it summarizes all the information contained in the data, and
2. knowing the individual values of $X_1, \dots, X_n$ gives no additional information on θ.

EG $X_1, \dots, X_n|\mu \sim$ iid normal μ, 1.

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sufficient for μ.

knowing $X_1, \cdots, X_n$ gives no additional information to estimate μ.

Definition. Let $X_1, \dots, X_n|\theta$ be iid random variables with $f(x|\theta)$ and $T = T(X_1, \dots, X_n)$ be a statistic. If the conditional distribution of $X_1, \dots, X_n|T, \theta$ depends only on T but not on θ, then T is a sufficient statistic.

dist. of $X_1, \cdots, X_n$ ~ depends on θ. dist. of $X_1, \cdots, X_n | T$ ~ independent of θ.

EG $X_1, \dots, X_n|\theta \sim$ iid Bernoulli θ. ← prob. of success.

Claim: $T(X_1, \cdots, X_n) = \sum_{i=1}^{n} X_i$ is sufficient for θ!

↖ # of successes in n trials.

$$P(X_1 = d_1, \cdots, X_n = d_n | T = t) = \frac{P(X_1 = d_1, \cdots, X_n = d_n, T = t)}{P(T = t)}$$

$T \sim$ Binomial n, θ.

$$= \begin{cases} \dfrac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \dfrac{1}{\binom{n}{t}}, & \sum d_i = t \\ 0, & \sum d_i \neq t \end{cases}$$

is independent of θ.

"Sufficient statistics"

$X_1, \cdots, X_n | \theta \sim f(x|\theta)$ (θ: unknown.)

Consider $T(X_1, \cdots, X_n)$ to estimate θ.

* T is considered "sufficient" for estimating θ, if it captures all the information about θ, (no need to examine $x_1, \cdots, x_n$ to estimate θ).

* Use Factorization thm. to find a sufficient statistic

$$f_n(x_1, \cdots, x_n|\theta) = u(x_1, \cdots, x_n) \times v\left[T(x_1, \cdots, x_n), \theta\right]$$

$\Rightarrow T(X_1, \cdots, X_n)$ is sufficient for θ.

Is there a simpler method for finding sufficient statistics?

Fisher – Neyman.

Theorem (Factorization Criterion). Let $X_1, \dots, X_n|\theta$ be iid random variables with $f(x|\theta)$ and $T = T(X_1, \dots, X_n)$ be a statistic. Then T is sufficient for θ if and only if

$f_n(x_1, \dots, x_n|\theta)$ can be factored as follows

likelihood function

$$f_n(x_1, \dots, x_n|\theta) = u(x_1, \dots, x_n) \times v[T(x_1, \dots, x_n), \theta],$$

u: depends on data; v: depends on a function of data and θ.

where u and v are nonnegative functions.

EG $X_1, \dots, X_n|\theta \sim$ iid Poisson θ

$$f_n(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta} \cdot \theta^{x_i}}{x_i!} = \frac{e^{-n\theta} \cdot \theta^{\sum x_i}}{\prod x_i!}$$

$$= \underbrace{\frac{1}{\prod x_i!}}_{u(x_1, \dots, x_n)} \times \underbrace{e^{-n\theta} \cdot \theta^{\sum x_i}}_{v[\sum x_i, \theta]} \Rightarrow \text{By Factorization thm.}$$

$T(X_1, \dots, X_n) = \sum X_i$ is sufficient for θ.

EG $X_1, \dots, X_n|\theta \sim$ iid uniform $[0, \theta]$

$$f_n(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \times \mathbb{1}(0 \le x_i \le \theta) = \frac{1}{\theta^n} \times \mathbb{1}(0 \le x_1, \dots, x_n \le \theta)$$

$$= \underbrace{\mathbb{1}(0 \le x_1, \dots, x_n)}_{u(x_1, \dots, x_n)} \times \underbrace{\frac{1}{\theta^n} \cdot \mathbb{1}(\max\{x_1, \dots, x_n\} \le \theta)}_{v[\max\{x_1, \dots, x_n\}, \theta]}$$

$\Rightarrow T = \max\{X_1, \dots, X_n\}$ is sufficient for θ

EG $X_1, \ldots, X_n|\theta \sim$ iid exponential θ $\quad f(x|\theta) = \theta e^{-\theta x}$

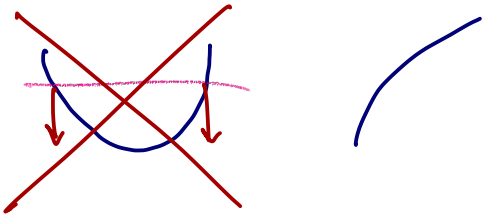
$$f_n(x_1,\cdots,x_n|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

$$= \underbrace{1}_{u} \times \underbrace{\theta^n e^{-\theta \sum_{i=1}^{n} x_i}}_{v\left[\sum x_i,\ \theta\right]}$$

By factorization thm., $T(X_1,\cdots,X_n) = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Is $h(T)$ also sufficient for θ when T is sufficient for θ?

Theorem Let T (1) be a sufficient statistic for θ, and h (2) be a 1-1 function. Then, $h(T)$ is a sufficient statistic for θ.



EG $X_1, \ldots, X_n|\theta \sim$ iid exponential θ $\quad$ rate $= \frac{1}{\text{average lifetime}}$

Is $\hat{\theta} = \text{MLE of } \theta = \frac{1}{\bar{X}_n} = \frac{n}{\sum_{i=1}^{n} X_i}$ sufficient for θ?

1. $\sum X_i$ is sufficient for θ.
2. $h(t) = \frac{n}{t}$, $t > 0$, is 1-1

$\Rightarrow h\left(\sum X_i\right) = \frac{n}{\sum X_i} = \hat{\theta}$ is sufficient for θ.

Ch7.8 Jointly sufficient statistics

What if your parameter is multidimensional $\boldsymbol{\theta} \in \Omega \subset \mathbb{R}^k$ for $k \geq 1$?

$(\theta_1, \theta_2, \cdots, \theta_k)$

$=$

EG $X_1, \ldots, X_n | \mu, \sigma^2 \sim$ iid normal μ, σ^2 $\Rightarrow$ $\theta = (\mu, \sigma^2)$, $k=2$.

EG $X_1, \ldots, X_n | a, b \sim$ iid uniform a, b $\Rightarrow$ $\theta = (a, b)$, $k=2$.

Definition. Statistics $(T_1, \ldots, T_k)$ are jointly sufficient for $\boldsymbol{\theta} \in \Omega \subset \mathbb{R}^k$, if the joint distribution of $(X_1, \ldots, X_n)$ given $(T_1, \ldots, T_k)$, $\boldsymbol{\theta}$ does not depend on $\boldsymbol{\theta}$.

Factorization Criterion for Jointly Sufficient Statistics.

$(T_1, \ldots, T_k)$ are jointly sufficient for θ if $f_n(x_1, \cdots, x_n | \theta)$ can be factored as follows

$x = (x_1, \cdots, x_n)$

$$f_n(x_1, \cdots, x_n | \theta) = u(x_1, \cdots, x_n) \times v\left[T_1(x), \cdots, T_k(x), \theta\right],$$

where u, v are nonnegative.

EG $X_1, \ldots, X_n | a, b \sim$ iid uniform $[a, b]$

$$f_n(x_1, \cdots, x_n | a, b) = \prod_{i=1}^{n} \frac{1}{b-a} \times \mathbb{1}(a \leq x_i \leq b).$$

$$= \frac{1}{(b-a)^n} \times \mathbb{1}(a \leq x_1, \cdots, x_n \leq b)$$

$$= 1 \times \frac{1}{(b-a)^n} \times \mathbb{1}(a \leq \min\{x_1, \cdots, x_n\} \leq \max\{x_1, \cdots, x_n\} \leq b).$$

$1 = u(x_1, \cdots, x_n)$

$v\left[\min\{x_1, \cdots x_n\}, \max\{x_1, \cdots x_n\}, a, b\right]$

$\Rightarrow$ By FT, $(\min\{X_1, \cdots X_n\}, \max\{X_1, \cdots, X_n\})$ is jointly sufficient for (a, b).

MLE of a ; MLE of b

EG $X_1, \ldots, X_n | \mu, \sigma^2$ iid normal μ, σ^2 $\rightsquigarrow f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$f_n(x_1,\cdots,x_n|\mu,\sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{\sqrt{2\pi}^n} \times \frac{1}{\sqrt{\sigma^2}^n} e^{-\frac{1}{2\sigma^2} \times \sum_{i=1}^n (x_i-\mu)^2}$$

$$\sum (x_i-\mu)^2 = \sum (x_i^2 - 2\mu x_i + \mu^2) = \sum x_i^2 - 2\mu \sum x_i + n\mu^2$$

$$= \frac{1}{\sqrt{2\pi}^n} \times \frac{1}{\sqrt{\sigma^2}^n} e^{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{2\mu}{2\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2}}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}^n}}_{u} \times \underbrace{\frac{1}{\sqrt{\sigma^2}^n} e^{\left[\sum x_i, \sum x_i^2\right]^*, \mu, \sigma^2}}_{v}$$

1. $\Rightarrow \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is jointly sufficient for (μ, σ^2).

Is $\left(\hat{\mu} = \bar{X}_n, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2\right)$ jointly sufficient for (μ, σ^2)?

$\bar{X}_n = \frac{\sum X_i}{n}$ // work $\frac{\sum X_i^2}{n} - \frac{(\sum X_i)^2}{n^2}$

2. $h(x,y) = \left(\frac{x}{n}, \frac{y}{n} - \frac{x^2}{n^2}\right)$ is 1-1,

$\Rightarrow (\hat{\mu}, \hat{\sigma}^2) = h\left(\sum X_i, \sum X_i^2\right)$ is jointly sufficient for (μ, σ^2).

Is there a set of jointly sufficient statistics that exists across all problems?

Definition. Let $X_1, \ldots, X_n$ be a random sample. Order the sample

$$\min\{x_1,\cdots,x_n\} = Y_1 \le Y_2 \le \cdots \le Y_{n-1} \le Y_n = \max\{x_1,\cdots,x_n\}.$$

$Y_1, \ldots, Y_n$ are called the order statistics.

Theorem. The order statistics are sufficient in any random samples.

$$\{Y_1,\cdots,Y_n\} = \{x_1,\cdots,x_n\}.$$

EG $X_1, X_2, X_3 | \mu, \sigma^2$ iid normal μ, σ^2
$x_1 = 1.1, x_2 = 1.6, x_3 = 0.3$

1. $\sum_{i=1}^{3} x_i = 3$, $\sum_{i=1}^{3} x_i^2 = 3.86$
2. $\hat{\mu} = \bar{x}_3 = 1$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{3}(x_i - \bar{x}_3)^2 = .43$
3. $y_1 = .3$, $y_2 = 1.1$ $y_3 = 1.6$

} three sets of sufficient statistics for (μ, σ^2)

To be useful, a sufficient statistic should be a simpler function of the data, capturing all the relevant information in the most concise form possible. In other words, we seek the simplest sufficient statistic! such as 1 and 2.

Definition. T is a minimal sufficient statistic if

1. T is sufficient for θ and

2. If T^* is sufficient, then there exists a function h such that $T = h(T^*)$.

EG $X_1, X_2, X_3 | \mu, \sigma^2$ iid normal μ, σ^2

Y_1, Y_2, Y_3 are sufficient, but not minimally sufficient for μ, σ^2.

$\nexists\ h$ s.t. $(Y_1, Y_2, Y_3) = h(\sum X_i, \sum X_i^2)$.

ex) $\exists\ h, h'$ s.t. $(\sum X_i, \sum X_i^2) = h(Y_1, Y_2, Y_3)$ or $h'(\hat{\mu}, \hat{\sigma}^2)$.

$\Rightarrow (\sum X_i, \sum X_i^2)$ is minimally sufficient.

How can we find minimal sufficient statistics?

Theorem. Let T be sufficient for θ.

✓ 1. The MLE $\hat{\theta}$ (or Bayes estimator) for θ depends on $X_1, \ldots, X_n$ only through T. ex) $\hat{\theta} = h(T)$.

✓ 2. $\hat{\theta}$ (or Bayes estimator) is sufficient, then it is minimally sufficient.

EG $X_1, \ldots, X_n | \mu, \sigma^2$ iid normal μ, σ^2

$\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

$\hat{\mu}$: MLE of μ; $\hat{\sigma}^2$: MLE of σ^2

✓ 1. $(\hat{\mu}, \hat{\sigma}^2)$ is a function of $(\sum X_i, \sum X_i^2)$ ← sufficient (see page 19)

✓ 2. $(\hat{\mu}, \hat{\sigma}^2)$ is sufficient (see page 18).

$\Rightarrow (\hat{\mu}, \hat{\sigma}^2)$ is minimally sufficient!

Ch7.9 Improving an estimator

In this section, we show how to improve upon an estimator using sufficient statistics.

Given: A random sample $X_1, \dots, X_n|\theta$ from $f(x_1|\theta)$. (θ: unknown parameter.)

Goal: Improve your estimator $\delta = \delta(\boldsymbol{X}) = \delta(X_1, \dots, X_n)$.

$\delta(x_1, \cdots, x_n)$: a function of $x_1, \cdots, x_n$, not necessarily sufficient statistics!

We first need to

1. pick a performance metric (ex) MSE and then
2. use vocabulary from decision theory.

Definition (Mean Squared Error). At given θ, the MSE of estimator δ is

$$R(\theta, \delta) = E\left[(\delta(X) - \theta)^2\right]$$

($R(\theta,\delta)$: risk of δ)

$\delta(X)$ is good if $R(\theta,\delta)$ is small across θ!

Definition.

1. Estimator δ_1 is at least as good as δ_2 if

$$R(\theta, \delta_1) \leq R(\theta, \delta_2), \quad \forall \theta \in \Omega.$$

2. Estimator δ_1 is admissible or dominates δ_2 if

- δ_1 is at least as good as δ_2, and
- $\exists\, \theta \in \Omega$ s.t $R(\theta, \delta_1) < R(\theta, \delta_2)$

(δ_1 is strictly better than δ_2 for some θ.)

We want to use admissible estimators!

Theorem (Rao-Blackwell). Let T be sufficient for θ, $\delta(\boldsymbol{X})$ be an estimator, and

$$\delta_0(T) = E[\delta(X) \mid T] \quad : \text{ a function of } T.$$

Then, δ_0 is at least as good as δ. If δ is not a function of a sufficient statistic, then δ_0 dominates δ.

"δ is not good"

EG $X_1, \dots, X_n | p \sim$ iid Bernoulli p. Let $\delta(\boldsymbol{x}) = x_1$.

(prob. success)

Sufficient statistic for p : $T = \sum_{i=1}^n X_i$ (Ch7.2)

Then, by R-B thm., $\delta_0 = E[X_1 \mid T]$ is at least as good as $\delta = X_1$, and δ_0 dominates δ because X_1 is not a function of $T = \Sigma X_i$

$$\delta_0 = E[X_1|T] = 1 \times P(X_1=1|T) + 0 \times P(X_1=0|T),$$

$$= \frac{P(X_1=1 \cap T=t)}{P(T=t)} = \frac{P(X_1=1 \cap \sum_{i=1}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)}$$

$$= \frac{P(X_1=1 \cap \sum_{i=2}^n X_i = t-1)}{P(\sum_{i=1}^n X_i = t)} = \frac{P(X_1=1)\, P(\sum_{i=2}^n X_i = t-1)}{P(\sum_{i=1}^n X_i = t)}$$

$\sum_{i=2}^n X_i \sim$ Binomial $(n-1, p)$

$X_1, \sum_{i=2}^n X_i$ are independent.

$\sum_{i=1}^n X_i \sim$ Binomial (n, p)

$$= \frac{p \cdot \binom{n-1}{t-1} p^{t-1} (1-p)^{(n-1)-(t-1)}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{\binom{n-1}{t-1}}{\binom{n}{t}} = \frac{\frac{(n-1)!}{(t-1)!\{(n-1)-(t-1)\}!}}{\frac{n!}{t!(n-t)!}}$$

$$= \frac{(n-1)!}{n!} \times \frac{t!}{(t-1)!} = \frac{t}{n} \quad (t = \Sigma X_i) \Rightarrow \delta_0 = \bar{X}_n.$$