What to do today (Nov 9, 2020)?

1. *Introduction*

2. *Probability and Distribution (Chp 1-3)*

3. **Essential Topics in Mathematical Statistics**

    3.1 Elementary Statistical Inferences (Chp 4)

    3.2 Consistency and Limiting Distributions (Chp 5)

    3.3 Maximum Likelihood Methods (Chp 6)

    - **3.3.1 Maximum Likelihood Estimation**
    - *3.3.2 Likelihood-Based Tests*
    - *3.3.3 EM Algorithm*

4. *Further Topics, Selected from Chp 7-11*

3.3.1 Maximum Likelihood Estimation (MLE): Procedure

Recall ... ...

Likelihood Function.

- Let the joint distribution (pmf, or pdf) of rvs $X_1, \ldots, X_n$ be $f(x_1, \ldots, x_n; \theta)$.
  When $x_1, \ldots, x_n$ are the observed values (realizations) of the rvs, the **likelihood function** of θ given the data is
  $$L(\theta \mid \text{ data }) = f(x_1, \ldots, x_n; \theta)$$
- **interpretation**: It's an overall measure on how likely the observed sample is the current set with the value of θ.
- Often $X_1, \ldots, X_n$ are iid observations (a random sample) from the population with distribution $f(x; \theta)$, $\theta \in \Omega$. If the observed values are $x_1, \ldots, x_n$, then the likelihood function is
  $$L(\theta \mid \text{ data }) = \prod_{i=1}^{n} f(x_i; \theta) = f(x_1; \theta) \ldots f(x_n; \theta).$$

Maximum Likelihood Estimator (MLE):

- The **MLE** $\hat{\theta}$ is the value of the population parameter θ that maximizes the likelihood function:

$$L(\hat{\theta} \mid \text{ data }) = \max_{\theta \in \Omega} L(\theta \mid \text{ data }).$$

- **interpretation**: The MLE $\hat{\theta}$ gives the parameter value that agrees most closely with the observed sample (the data).

- Often used **procedures**:

(1) to maximize $\log L(\theta)$
(2) to obtain the solution to the *likelihood estimating equation*
$\partial \log L(\theta) / \partial \theta = 0$

Example 3.7 Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli distn $B(1, \theta)$. What is the MLE of θ?

Example 3.8 Let iid $X_1, \ldots, X_n \sim f(x; \theta) = e^{-(x-\theta)} / (1 + e^{-(x-\theta)})^2$ for $x \in (-\infty, \infty)$ and $\theta \in (-\infty, \infty)$ (*Logistic Distribution*). What is the MLE of θ?

3.3.1 Maximum Likelihood Estimation (MLE): Rationale

Assumptions. (Regularity Conditions) Consider $\{f(x;\theta) : \theta \in \Omega\}$.

(R0) If $\theta \neq \theta^*$, $f(\cdot;\theta) \neq f(\cdot;\theta^*)$.

(R1) $\{f(x;\theta) : \theta \in \Omega\}$ have common support.

(R2) θ_0 is an interior point in Ω.

Theorem. Consider rv $X \sim f(x;\theta)$ for $\theta \in \Omega$ with a random sample $X_1, \ldots, X_n$. If θ_0 is the true value of θ, provided (R0)-(R2), for $\theta \in \Omega$

$$\lim_{n\to\infty} P_{\theta_0}\big[L(\theta_0 \mid X_1, \ldots, X_n) > L(\theta \mid X_1, \ldots, X_n)\big] = 1.$$

Definition. (MLE) With the random sample $X_1, \ldots, X_n$, $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ is the MLE if $\widehat{\theta} = argmax_{\theta\in\Omega} L(\theta \mid X_1, \ldots, X_n)$.

3.3.1 Maximum Likelihood Estimation (MLE): Properties

Let iid $X_1, \dots, X_n \sim f(x;\theta), \theta \in \Omega$.

Theorem. (Invariance) Let iid $X_1, \dots, X_n \sim f(x;\theta), \theta \in \Omega$. If $\widehat{\theta}$ is the MLE of θ, $\widehat{\eta} = g(\widehat{\theta})$ is the MLE of $\eta = g(\theta)$.
Proof: Note that

$$\max_{\eta} L(\eta|data) = \max_{\theta:\eta=g(\theta)} L(g(\theta)|data) = L(g(\widehat{\theta})|data).$$

Theorem. Provided (R0)-(R2) and $f(x;\theta)$ is differentiable wrt $\theta \in \Omega$, if θ_0 is the true value, the likelihood equation $\partial L(\theta) \big/ \partial\theta = 0$ or $\partial \log L(\theta) \big/ \partial\theta = 0$ has a solution $\widehat{\theta}_n$ such that $\widehat{\theta}_n \xrightarrow{P} \theta_0$.

$\Longrightarrow$ If the MLE of θ is the solution, it is *consistent*.

3.3.1 Maximum Likelihood Estimation (MLE): Properties

Assumptions. (Additional Regularity Conditions) Consider $\{f(x;\theta) : \theta \in \Omega\}$.

(R3) $f(x;\theta)$ is twice differentiable wrt θ.

(R4) $E\left[\partial \log f(X;\theta)/\partial\theta\right]$ and $E\left[\partial^2 \log f(X;\theta)/\partial\theta^2\right]$ exist.

(R5) $f(x;\theta)$ is three times differentiable wrt θ.
$\left|\partial^3 \log f(X;\theta)/\partial\theta^3\right| \leq M(x)$ for $\theta \in (\theta_0 - c, \theta_0 + c)$ and all x in the support of X, and $E_{\theta_0}\left[M(X)\right] < \infty$.

3.3.1 Maximum Likelihood Estimation (MLE): Properties

Definition. (Fisher Information) The Fisher information is $FI(\theta) = E\left[\left(\partial \log f(X;\theta)/\partial\theta\right)^2\right]$, provided the expectation exists.

Note that

$$FI(\theta) = Var\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right) = -E\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right].$$

Theorem (Asymptotic Normality) Provided (R0)-(R5) and $0 < FI(\theta_0) < \infty$, the solution $\widehat{\theta}_n$ to the likelihood equation $\partial L(\theta)/\partial\theta = 0$ or $\partial \log L(\theta)/\partial\theta = 0$ satisfies $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{D} N(0, FI(\theta_0)^{-1})$.

Proof: Expand $\partial \log L(\theta)/\partial\theta = l^{'}(\theta)$ into the Taylor series of order 2 about θ_0 and evaluate it at $\widehat{\theta}_n$:

$$l^{'}(\widehat{\theta}_n) = l^{'}(\theta_0) + (\widehat{\theta}_n - \theta_0)l^{''}(\theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^2 l^{'''}(\theta_n^*),$$

θ_n^* in between θ_0 and $\widehat{\theta}_n$. Note that $l^{'}(\widehat{\theta}_n) = 0$,

$$\frac{1}{\sqrt{n}} l^{'}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta_0)}{\partial\theta} \xrightarrow{D} N(0, FI(\theta_0))$$

by CLT, and

$$-\frac{1}{n} l^{''}(\theta_0) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta_0)}{\partial\theta^2} \xrightarrow{P} FI(\theta_0).$$

Further $\left| -\frac{1}{n} l^{'''}(\theta_n^*) \right| \leq \frac{1}{n}\sum_{i=1}^{n} M(X_i)$, and thus $l^{'''}(\theta_n^*)/n$ is bounded in probability by (R5). Combining the results yields the theorem.

3.3.1 Maximum Likelihood Estimation (MLE): Cramer-Rao Lower Bound and Efficiency

Theorem. (Cramer-Rao Lower Bound) Let iid $X_1, \ldots, X_n \sim f(x;\theta)$ for $\theta \in \Omega$. Assume (R0)-(R4). Let $Y = u(X_1, \ldots, X_n)$ be a statistic and $E(Y) = k(\theta)$. Then

$$Var(Y) \geq \frac{\left(k'(\theta)\right)^2}{nFI(\theta)}.$$

$\Longrightarrow Var(Y) \geq \frac{1}{nFI(\theta)}$ if Y is an unbiased estimator of θ.

Definition. An unbiased estimator Y with a random sample of size n is called **efficient** if $Var(Y) = \frac{1}{nFI(\theta)}$.

$\Longrightarrow$ The MLE $\widehat{\theta}$ is *asymptotically efficient*.

Example 3.9 (Beta Distribution) Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta) = \theta x^{\theta-1}$ for $0 < x < 1$ and $\theta \in \Omega = (0, \infty)$.

3.3.1 Maximum Likelihood Estimation (MLE): Multiparameter Case

Likelihood Function.

- Let the joint distribution (pmf, or pdf) of rvs $X_1, \ldots, X_n$ be $f(x_1, \ldots, x_n; \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^{'} \in \Omega \subseteq \mathcal{R}^K$. When $x_1, \ldots, x_n$ are the observed values (realizations) of the rvs, the **likelihood function** of $\boldsymbol{\theta}$ given the data is

$$L(\boldsymbol{\theta} \mid \text{ data }) = f(x_1, \ldots, x_n; \boldsymbol{\theta})$$

- **interpretation**: It's an overall measure on how likely the observed sample is the current set with the value of $\boldsymbol{\theta}$.

- Often $X_1, \ldots, X_n$ are iid observations (a random sample) from the population with distribution $f(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Omega$. If the observed values are $x_1, \ldots, x_n$, then the likelihood function is

$$L(\boldsymbol{\theta} \mid \text{ data }) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = f(x_1; \boldsymbol{\theta}) \ldots f(x_n; \boldsymbol{\theta}).$$

Maximum Likelihood Estimator (MLE):

- The **MLE** $\widehat{\boldsymbol{\theta}}$ is the value of the population parameter $\boldsymbol{\theta}$ that maximizes the likelihood function:

$$L(\widehat{\boldsymbol{\theta}} \mid \text{ data }) = \max_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta} \mid \text{ data }).$$

- **interpretation**: The MLE $\widehat{\boldsymbol{\theta}}$ gives the parameter value that agrees most closely with the observed sample (the data).

- Often used **procedures**:

(1) to maximize $\log L(\boldsymbol{\theta})$.
(2) to obtain the solution to the *likelihood estimating equation* $\bigtriangledown \log L(\boldsymbol{\theta}) = \partial \log L(\boldsymbol{\theta}) \big/ \partial \boldsymbol{\theta} = \mathbf{0}$. (The gradian $\bigtriangledown g(\mathbf{u}) = \frac{\partial g(\mathbf{u})}{\partial \mathbf{u}}$.)

Example 3.10 Suppose iid rvs $X_1, \ldots, X_n$ with the pdf

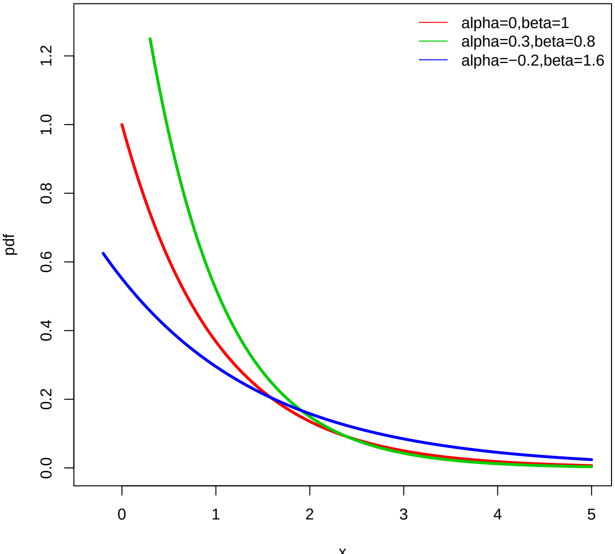
$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta}\exp\{-(\frac{x-\alpha}{\beta})\}, & x \geq \alpha \\ 0, & \textit{elsewhere} \end{cases}$$

Derive the MLE of the parameter $\boldsymbol{\theta} = (\alpha, \beta)'$ for $\alpha \in (-\infty, \infty)$ and $\beta \in (0, \infty)$.

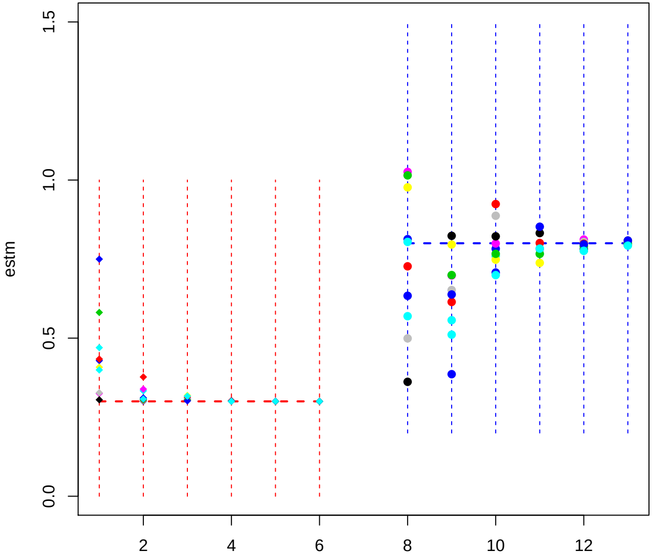
Exponential distn: pdf

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta}\exp\{-(\frac{x-\alpha}{\beta})\}, & x \geq \alpha \\ 0, & elsewhere \end{cases}$$

- $X = \beta T + \alpha$ with $T \sim NE(1)$



Generate iid $x_1, \ldots, x_n$ from the exponential-distn with $\alpha = 0.3$ and $\beta = 0.8$: $n = 5, 5^2, 5^3, 5^4, 5^5, 5^6$ and repeat each setting 10 times to evaluate the MLE $\hat{\alpha}$ and $\hat{\beta}$



3.3.1 Maximum Likelihood Estimation (MLE): Multiparameter Case

Consider rv $X \sim f(x; \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$.

Definition. (Fisher Information) The Fisher information is

$$\mathrm{FI}(\boldsymbol{\theta}) = E\Big[\Big(\bigtriangledown \log f(X; \boldsymbol{\theta})\Big)\Big(\bigtriangledown \log f(X; \boldsymbol{\theta})\Big)'\Big],$$

provided the expectation exists. $\mathrm{FI}(\boldsymbol{\theta})$ is $K \times K$, nonnegative definite.

Note that $E\big(\bigtriangledown \log f(X; \boldsymbol{\theta})\big) = 0$ and then

$$\mathrm{FI}(\boldsymbol{\theta}) = Var\Big(\bigtriangledown \log f(X; \boldsymbol{\theta})\Big) = -E\Big[\frac{\partial^2 \log f(X; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\Big].$$

The (j, k)th entry of $\mathrm{FI}(\boldsymbol{\theta})$ for $j, k = 1, \ldots, K$:

$$\mathrm{FI}_{jk} = E\Big[\Big(\frac{\partial \log f(X; \boldsymbol{\theta})}{\partial \theta_j}\Big)\Big(\frac{\partial \log f(X; \boldsymbol{\theta})}{\partial \theta_k}\Big)\Big] = Cov\Big(\frac{\partial \log f(X; \boldsymbol{\theta})}{\partial \theta_j}, \frac{\partial \log f(X; \boldsymbol{\theta})}{\partial \theta_k}\Big)$$

Suppose $X_1, \ldots, X_n \sim f(x; \boldsymbol{\theta})$ iid with $\boldsymbol{\theta} \in \Omega \subseteq \mathcal{R}^K$.

Theorem (Asymptotic Properties) Provided (R0)-(R5) in the multiparameter case hold. Then

1. The likelihood equation $\partial \log L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}$ has a solution $\widehat{\boldsymbol{\theta}}_n$ such that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}$.
2. For such $\widehat{\boldsymbol{\theta}}_n$, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} MN(\mathbf{0}, \mathrm{FI}(\boldsymbol{\theta})^{-1})$.

Corollary For $j = 1, \ldots, K$, the jth component of $\widehat{\boldsymbol{\theta}}_n$ satisfies

$$\sqrt{n}(\widehat{\theta}_{n,j} - \theta_j) \xrightarrow{D} N(0, [\mathrm{FI}(\boldsymbol{\theta})^{-1}]_{jj}).$$

Example 3.11 Consider an experiment with 3 different types of outcome and the corresponding probabilities $\theta_1, \theta_2, \theta_3$. ($\sum \theta_j = 1$). Let the 3 componets of $\mathbf{X} = (X_1, X_2, X_3)$ be the indicators of the 1st, 2nd, 3rd types: $\mathbf{X} \sim$ *trinomial* distn:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3} = \theta_1^{x_1} \theta_2^{x_2} (1 - \theta_1 - \theta_2)^{1 - x_1 - x_2}.$$

Fisher Information Matrix: $\bigtriangledown \log f(\mathbf{x}; \boldsymbol{\theta}) = \left(\frac{x_1}{\theta_1} - \frac{x_3}{\theta_3}, \frac{x_2}{\theta_2} - \frac{x_3}{\theta_3} \right)'$

$$\frac{\partial^2 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_1^2} = -\frac{x_1}{\theta_1^2} - \frac{x_3}{\theta_3^2}, \quad \frac{\partial^2 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_2^2} = -\frac{x_2}{\theta_2^2} - \frac{x_3}{\theta_3^2}, \quad \frac{\partial^2 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} = -\frac{x_3}{\theta_3^2}$$

The entries of $\mathrm{FI}(\boldsymbol{\theta})$ are

$$\mathrm{FI}_{11} = \frac{1}{\theta_1} + \frac{1}{\theta_3}, \quad \mathrm{FI}_{12} = \mathrm{FI}_{21} = \frac{1}{\theta_3}, \quad \mathrm{FI}_{22} = \frac{1}{\theta_2} + \frac{1}{\theta_3}.$$

MLE with $\mathbf{x}_1, \ldots, \mathbf{x}_n$*:* $L(\boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^n \theta_1^{x_{1i}} \theta_2^{x_{2i}} \theta_3^{x_{3i}}$.

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^n x_{1i} \log \theta_1 + \sum_{i=1}^n x_{2i} \log \theta_2 + \sum_{i=1}^n x_{3i} \log \theta_3$$

For $h = 1, 2$, $\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_h} = \frac{\sum_{i=1}^n x_{hi}}{\theta_h} - \frac{\sum_{i=1}^n x_{3i}}{\theta_3} = 0 \Longrightarrow \widehat{\theta}_j = \frac{\sum_{i=1}^n x_{ji}}{n}$ for $j = 1, 2, 3$

What will we do next week?

1. *Introduction*

2. *Probability and Distribution (Chp 1-3)*

3. **Essential Topics in Mathematical Statistics (Chp 4-6)**

   3.1 Elementary Statistical Inferences (Chp 4)
   3.2 Consistency and Limiting Distributions (Chp 5)
   3.3 Maximum Likelihood Methods (Chp 6)

   - *3.3.1 Maximum Likelihood Estimation*
   - **3.3.2 Likelihood-Based Tests**
   - **3.3.3 EM Algorithm**

4. *Further Topics, Selected from Chp 7-11*