STAT 514 Lecture Notes

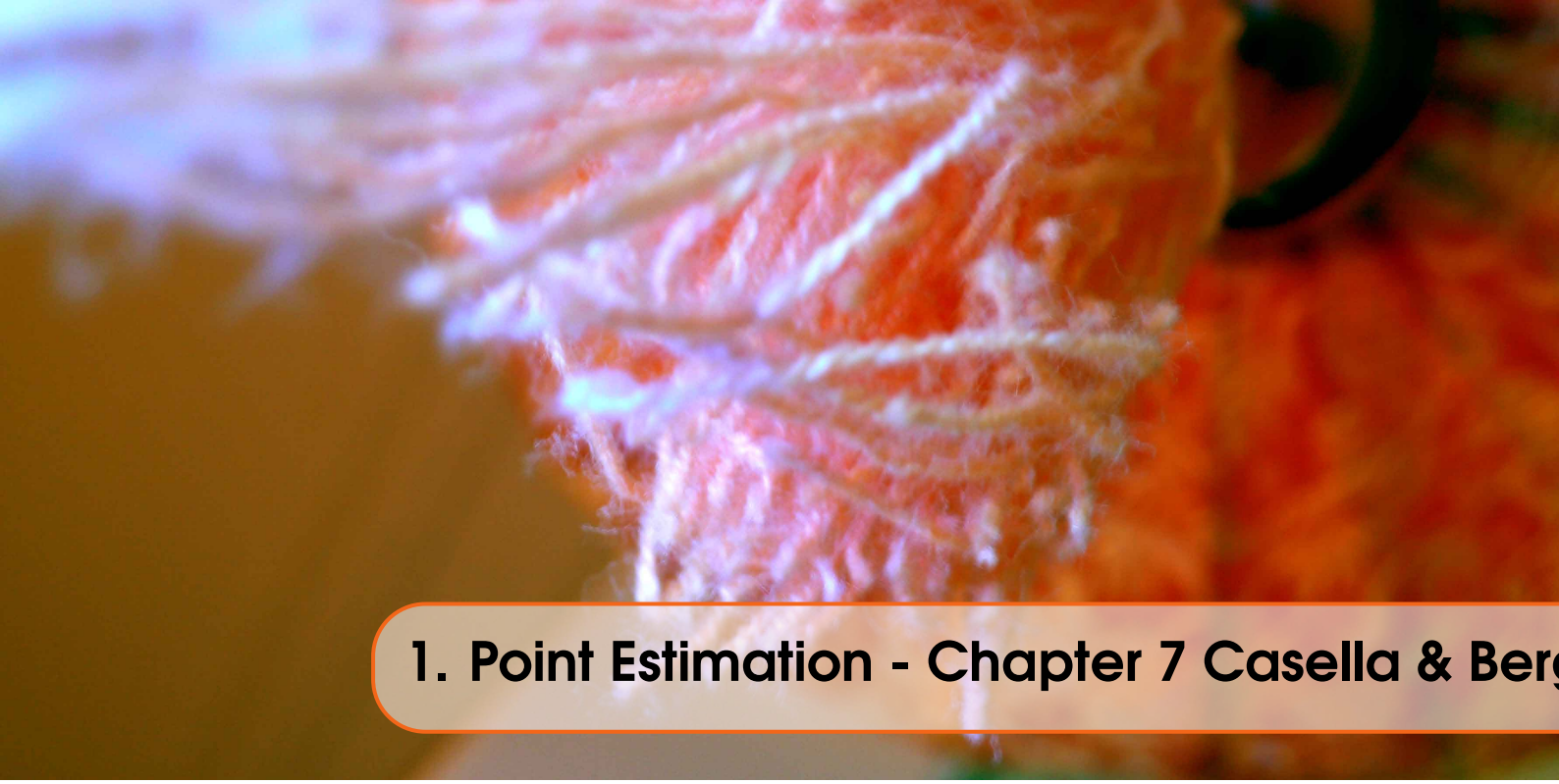
Dr. David Hunter

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

First printing, March 2013

Contents



1. Point Estimation - Chapter 7 Casella & Ber

1.1 Introduction

In the simplest case, we have n observations of data that we believe follow the same distribution.

$$X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$$

where $f_\theta(x)$ is a density function involving a parameter θ. Our goal is to learn something about θ, which could be real or vector valued.

Definition 1.1.1 — Estimator. An *estimator* of θ is any function $W(X_1, \ldots, X_n)$ of the data. That is, an estimator is a *statistic*.

Note:

1. $W(\boldsymbol{X})$ may not depend on θ.
2. $W(\boldsymbol{X})$ should resemble or "be close" to θ.
3. An estimator is *random*.
4. $W(X_1, \ldots, X_n)$ is the estimator, $W(x_1, \ldots, x_n)$ is the fixed estimate.

■ **Example 1.1** Suppose we have n observations from an exponential distribution,

$$X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x) = \frac{1}{\theta} \exp\left\{-\frac{x}{\theta}\right\} \mathbb{1}\{x > 0\}$$

for some $\theta > 0$. The **likelihood function** is equivalent to the joint density function, expressed as a function of θ rather than the data:

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \exp\left\{-\frac{x}{\theta}\right\} = \frac{1}{\theta^n} \exp\left\{-\frac{1}{\theta} \sum_{i=1}^{n} x_i\right\}$$

This function represents the *likelihood* of observing the data we observed assuming the parameter was a particular value of θ. If we can maximize this function, we can determine the $\hat{\theta}$ for which the likelihood of observing $\boldsymbol{X}$ was the highest. This might tell us something about the true value of θ.

To maximize $L(\theta)$, we want to take the derivative, set it equal to 0, and solve for θ. However, in many cases taking the derivative of the likelihood function will be very hard, if not impossible.

We can use the fact that taking the logarithm does not change the location of extrema. The **log-likelihood function** in this case is

$$\ell(\theta) = \log L(\theta) = -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

Take the derivative with respect to the parameter and set equal to 0:

$$\begin{aligned} \ell'(\theta) &= -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \overset{\text{set}}{=} 0 \\ \hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Here $\hat{\theta}$ is an estimator (the sample mean). Since it maximizes $L(\theta)$, we call it the **maximum likelihood estimator** (MLE). ∎

1.2 Methods of Finding Estimators

Setup: $x_1,\ldots,x_n \overset{iid}{\sim} f_\theta, \theta \in \Theta$
Goal: Estimate θ, or some function $\tau(\theta)$

Definition 1.2.1 — Point Estimator. A point estimator $T(x_1,\ldots,x_n)$ is a function of the random sample. T should not depend on θ.

R Estimator is defined above, while an estimate is a fixed value. ($T(x_1 = x_1,\ldots,x_n = x_n)$)

1.2.1 Method of Moments

This method is one of the oldest, simplest estimators developed by Karl Pearson.

Construct: $\begin{cases} \frac{1}{n}\sum X_i = E_\theta(x) = \int_\theta x f_\theta(x)dx \\ \frac{1}{n}\sum X_i^2 = E_\theta(x^2) = \int_\theta x^2 f_\theta(x)dx \\ \vdots \end{cases}$

So, $\hat{\theta}_M M$ depends only on $(\frac{1}{n}\sum X_i = E_\theta(x),\ldots,\frac{1}{n}\sum X_i = E_\theta(x^k))$.

■ **Example 1.2** $x_1,\ldots,x_n \overset{iid}{\sim} N(\mu,\sigma^2)$
μ,σ^2 unknown: $\Theta = \mathbb{R}x\mathbb{R}^+$

$$\begin{aligned} \frac{1}{n}\sum X_i = \mu &\implies \hat{\mu_{MM}} = \frac{1}{n}\sum X_i \triangleq \bar{X} \\ \frac{1}{n}\sum X_i^2 = \sigma^2 + \mu^2 &\implies \hat{\sigma_M M^2} = \frac{1}{n}\sum (X_i - \bar{X})^2 \end{aligned}$$

∎

■ **Example 1.3** $x_1,\ldots,x_n \overset{iid}{\sim} \text{Bin}(k,p)$ where k,p are unknown: $\Theta : \mathbb{N}$ x $[0,1]$.

$$\begin{aligned} &\frac{1}{n}\sum X_i = kp \triangleq \bar{X} \\ &\frac{1}{n}\sum X_i^2 = kp(1-p) + k^2p^2 = \bar{X}(1-p) + \bar{X}^2 \end{aligned}$$

∎

R Read about Satterthwait approximation

1.2.2 Maximum Likelihood Estimation

MLE makes data the most likely under the model specified. The quality of MLE depends on the quality of the model. Unlike Method of Moments, MLE does not lie outside of Θ.

Invariance Principle: if $\hat{\theta}$ is MLE for θ, then $\tau(\hat{\theta})$ is MLE for $\tau(\theta)$.

Roughly speaking, MLEs enjoy a nice property called **asymptotic efficiency**. They acheive the smallest possible varience in large sample. Similar to Lower Bound of Cramer-Rao as n -> infinity.

Definition 1.2.2 — Score Function. Suppose that throughout $X_1,\ldots,X_n \sim f_\theta(x)$ the **score function** is $\sum \frac{\delta}{\delta\theta} \log f_\theta(x_i)$.

Take a first-order Taylor expansion of $\ell(\theta)$ around true value (θ).

$$
\begin{aligned}
\ell'(\hat{\theta}) &\simeq \ell'(\theta_0) + (\hat{\theta}-\theta_0)\ell''(\theta_0) \\
\hat{\theta}-\theta_0 &\simeq [\ell''(\theta_0)^{-}1][\ell'(\hat{\theta})-\ell'(\theta_0)] \\
&= [\frac{-1}{n}\ell''(\theta_0)^{-1}][\frac{1}{n}\ell'(\theta_0)] \\
\frac{(\hat{\theta}-\theta_0)}{\sqrt{n}} &\simeq \frac{\frac{-1}{\sqrt{n}}\ell'(\theta_0)}{\frac{-1}{n}\ell''(\theta_0)} \\
\frac{1}{n}\ell'(\theta_0) &= \frac{1}{n}\sum \frac{\delta}{\delta\theta}\log f_\theta(x) \\
\frac{-1}{n}\ell''(\theta_0) &= \frac{1}{n}\sum \frac{\delta^2}{\delta\theta^2}\log f_\theta(x)|_{\theta=\theta_0}
\end{aligned}
$$

Thus, $\sqrt{n}(\hat{\theta}-\theta_0) \sim N(0,[I(\theta_0)]^{-1})$

Finding the Maxima: $\theta \in \mathbb{R}^k$
Suppose L is differentialbe, then the only possible candidtes for $\hat{\theta_{MLE}}$ are the stationary points of L which satisfy $\frac{\delta L}{\delta\theta_i} = 0, i = 1,\ldots,k$, which is a necessary condition for optimality.
Suppose L is twice differentiable,

$$
\mathrm{H} \triangleq \begin{bmatrix} \frac{\delta^2 L}{\delta\theta_1^2} & \frac{\delta^2 L}{\delta\theta_1\delta\theta_2} & \cdots & \frac{\delta^2 L}{\delta\theta_1\delta\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta^2 L}{\delta\theta_k\theta_1} & \cdots & \cdots & \frac{\delta^2 L}{\delta\theta_k^2} \end{bmatrix}
$$

If $H|_{\theta=\hat{\theta_{ML}}}$ is negative definite, then $\hat{thetaML}$ is the MLE.

Definition 1.2.3 — Negative Definite. A matrix, A_{nxn} is called **negative definite** if for any $(a_1,a_2,dots,a_n) \in \mathbb{R}^m$, $\sum_{i,j=1}^{n} a_i a_j A_{ij} < 0$. Equivalent characterization is the eigenvalues of A are negative.

R More reading: C&B sections 3.4, 7.2.3.

R MLE and MM should match on: Bernoulli, Binomial, Normal, and Poisson distrabutions.

1.2.3 Bayes' Estimation

Theorem 1.2.1 — Bayes' Theorem.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

LHS: posterior density of the parameters given data. This is what we want to know. The RHS includes

i) likelihood
ii) prior density
iii) marginal density of x

Proposition 1.2.2 — Dave's Bayes Way. posterior $\propto$ prior*likelihood

R Bayesian estimators regularize MLE.

Definition 1.2.4 — Conjugate Family. If the prior is chosen from a family of distributions in such a way that the posterior is also a member, then we call this family of priors a **conjugate family**.

Likelihood	Conjugate Prior
Binom(p)	Beta(α,β)
Poisson(λ)	Gamma(α,β)
Exp(λ)	Gamma(α,β)
Normal(μ,σ^2)	Gamma(α,β)

1.3 Evaluating Estimation

1.3.1 Mean Squared Error

R Read Casella & Berger Chapter 7.3 - Methods of Evaluating Estimation

Definition 1.3.1 — Mean Squared Error. If $W(\boldsymbol{X})$ is an estimator of θ, then the **mean squared error** (MSE) is defined as

$$E_\theta\left[(W(\boldsymbol{X})-\theta)^2\right].$$

Definition 1.3.2 — Unbiased estimator. If $W(\boldsymbol{X})$ is an estimator of θ, we say that $W(\boldsymbol{X})$ is **unbiased** if

$$E_\theta[W(\boldsymbol{X})] = \theta \quad \forall\theta.$$

Furthermore, the **bias** of $W(\boldsymbol{X})$ is

$$E_\theta[W(\boldsymbol{X})] - \theta.$$

■ **Example 1.4** For $\theta > 0$, let

$$X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x) = \theta x^{-2} \mathbb{1}\{x > \theta\}$$

Find the MLE of θ.

$$L(\theta) = \theta^n \prod_{i=1}^{n} x_i^{-2} \prod_{i=1}^{n} \mathbb{1}\{x > \theta\}$$

$$\hat{\theta} = \text{minimum of } x_i$$

■

Theorem 1.3.1 $\text{MSE}(W) = \text{Bias}^2 + \text{Var}(W)$

Proof.

$$\begin{aligned} E[(W(\boldsymbol{X}) - \theta)^2] &= E[(W - E[W] + E[W] - \theta)^2] \\ &= E[(W - E[W])^2] + E[(E[W] - \theta)^2] + 2E[(W - E[W])(E[W] - \theta)] \\ &= \text{Var}(W) + \text{Bias}^2(W) + 0 \end{aligned}$$

■

■ **Example 1.5** $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 is known. $\hat{\mu}_{ML} = \bar{X}$; and $\hat{\mu} = X_1$

$$\begin{aligned} \text{MSE}(\hat{\mu}_{ML}) &= [E(\hat{\mu}_{ML} - \mu)]^2 + \text{Var}(\bar{X}) \\ &= E(\bar{X}) - \mu + \frac{\sigma^2}{n} \\ &= \frac{\sigma^2}{n} \end{aligned} \qquad \begin{aligned} \text{MSE}(\hat{\mu}) &= [E(\hat{\mu} - \mu)]^2 + \text{Var}(X_1) \\ &= E(X_1) - \mu + \sigma^2 \\ &= \sigma^2 \end{aligned}$$

So, for n > 1, $\text{MSE}(\hat{\mu}_{ML})$ is better than $\text{MSE}(\hat{\mu})$. ■

1.3.2 Best Unbiased Estimator

What does "best" mean? Answer: Minimum variance.

While we can NOT find estimators that have smallest MSE for and θ, we CAN find estimators within the class of unbiased estimators that achieve smallest MSE for all $\theta \in \Theta$.

Definition 1.3.3 — Best Unbiased Estimator. An estimator W* is a **best unbiased estimator*** of $\tau(\theta)$ if it satisties

$$E_\theta(W*) = \tau(\theta)$$

for all θ and, for any other estimator W with

$$E_\theta(W) = \tau(\theta)$$

we have

$$\text{Var}_\theta(W*) \leq \text{Var}_\theta(W)$$

for all θ. W* is also called a **uniform minimum variance unbiased estimator** (UMVUE) of $\tau(\theta)$.

Main Result: Under some assumptions we can establish a lower bound on Var(W(X)). The question of how small this variance could be is answered by Cramer-Rao Inequality.

R Spoilers: We will show that there are two main route for finding the UMVUE.

1. Check if estimator attains the Cramer-Rao Lower Bound.
2. Find a complete and sufficient statistic then a function of it that is unbiased.
   (a) Check to see if density if part of the exponential family. T is complete and sufficient.
   (b) Use factorization theorem to find a sufficient statistic then the definition of completeness and Lehmann-Sheffe to get UMVUE

But first, we need the following:

Theorem 1.3.2 — Cauchy-Scwarz Inequality.

$$|| < x,y > ||^2 \leq || < x,x > || * || < y,y > ||$$

In terms of E(X) if A & B have $\mu = 0$:

$$(E(AB))^2 \leq E(A^2) * E(B^2)$$

Or in terms of covariance:

$$Cov^2(AB) \leq Var(A) * Var(B)$$

Proof. Let $D = B - \frac{E(AB)}{E(A^2)}A$, given $D^2 \geq 0$.

$$\begin{aligned} E(D^2) &= E(B^2 - 2(B)(\frac{E(AB)}{E(A^2)}A) + (\frac{E(AB)}{E(A^2)}A)^2) \\ &= E(B^2) - 2\left(\frac{E(AB)^2}{E(A^2)}\right) + \frac{E(AB)}{E(A^2)}E(A^2) \\ &= E(B^2) - \frac{E(AB)^2}{E(A^2)} \geq 0 \end{aligned}$$

■

Theorem 1.3.3 — Cramer-Rao Inequality. Also: Information Inequality.

$$\text{Var}(W(\underline{X})) \geq \frac{(\Psi'(\theta))^2}{I(\theta)}$$

where, $\Psi(\theta) = E_\theta(W(\underline{X}))$
and, the Fisher/Expected information, $I(\theta) = E\left((\frac{\delta}{\delta\theta}\log f_\theta(\underline{X}))^2\right)$.

Proof. Follows from Cauchy-Schwarz Inequality

$$Cov(W(\underline{x}), \frac{\delta}{\delta\theta}\log f_\theta(\underline{X})))^2 \leq \text{Var}(W(\underline{X})) * \text{Var}(\frac{\delta}{\delta\theta}\log f_\theta(\underline{X}))$$

■

Definition 1.3.4 — Regular. A family, $\mathscr{F} = \{f_\theta : \theta \in \Theta\}$ is called **regular** if:

(I) The set $A = \{x : f_\theta(x) > 0\}$ does not depend on θ. Forall $x \in A, \theta \in \Theta$, $\frac{\delta}{\delta\theta}\log(f_\theta(x))$ exists and is finite.

(II) For any estimator W such that $E_\theta|W| < \infty$ for all $\theta \in \Theta$, the operations of integration and differentiation by θ can be interchanged when the r.h.s is finite.

Assumptions:

1. $\mathrm{I}(\theta) = \mathrm{E}_\theta\left((\frac{\delta}{\delta\theta}\log f_\theta(\underline{X}))^2\right) = \mathrm{Var}\left((\frac{\delta}{\delta\theta}\log f_\theta(\underline{X})^2)\right)$ is well defined and $\mathrm{I}(\theta) > 0$.
2. $E\left((\frac{\delta}{\delta\theta}\log f_\theta(\underline{X}))^2\right) = 0$. Thus, $\mathrm{I}(\theta)$ is just varience.
3. $E\left((W(\underline{X})\frac{\delta}{\delta\theta}\log f_\theta(\underline{X}))^2\right) = \Psi'(\theta)$

So,

$Cov(W(\underline{x}), \frac{\delta}{\delta\theta}\log f_\theta(\underline{X})))^2 \leq \mathrm{Var}(W(\underline{X})) * \mathrm{Var}(\frac{\delta}{\delta\theta}\log f_\theta(\underline{X})) \simeq \mathrm{Var}(W(\underline{X})) \geq \frac{(\Psi'(\theta))^2}{I(\theta)}$

R

(i) Suppose W(x) is an unbiased estimator of θ. Then $(\Psi'(\theta))^2 = 1$ and the bound has nothing to do with the estimator anymore, but only deals with f_θ.

(ii) Suppose $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta$, then $I(\theta) = nI(\theta)$ where the latter information is related to only a single variable, X.

(iii) This result also holds for discrete distributions (as long as the summation and differentiation are interchangeable).

(iv) The CR bound might be stictly smaller than the varience of any estimator.

(v) Under additional assumptions on f_θ,

$$I(\theta) = -E\left[\frac{\delta^2}{\delta\theta^2}\log(f_\theta(x))\right]$$

Exercise 1.1 Let $\underline{X} \sim Poi(\theta)$, Y=$\sum x_i$, and Y $\sim Poi(n\theta)$. What is $I(\theta)$?

$$\begin{aligned}
I(\theta) &= E\left((\frac{\delta}{\delta\theta}\log f_\theta(\underline{X})^2)\right) \\
&= E\left((\frac{\delta}{\delta\theta}\log\prod\frac{\theta^x e^{-\theta}}{x!})^2)\right) \\
&= E\left((\frac{\delta}{\delta\theta}\log\frac{\theta^{\sum x_i}e^{-\theta n}}{\sum x_i!})^2\right) \\
&= E\left(\frac{\delta}{\delta\theta}(-n\theta + \sum x_i\log\theta - \sum\log x_i!)^2\right) \\
&= E\left((-n + \frac{\sum x_i}{\theta})^2\right) \\
&= E\left(n^2 - 2(\frac{\sum x_i}{\theta})(n) + (\frac{\sum x_i}{\theta})^2\right) \\
&= n^2 - \frac{2n * E(\sum x_i)}{\theta} + \frac{E((\sum x_i)^2)}{\theta^2} \\
&= n^2 - \frac{2n * E(Y)}{\theta} + \frac{E(Y)^2}{\theta} \\
&= \frac{n}{\theta}
\end{aligned}$$

Note: I(θ) is equal to information in the whole sample, but sometimes it's just one sample (based on context).
If we assume (as in any exponential family):

$$E_\theta\left(\frac{\delta^2}{\delta\theta^2}\log f_\theta(\underline{x})\right) = \frac{\delta^2}{\delta\theta^2}\int \log f_\theta(\underline{x}) f_\theta(\underline{x}) dx$$

then, the observed information is

$$I(\theta) = -E_\theta\left(\frac{\delta^2}{\delta\theta^2}\log f_\theta(\underline{x})\right)$$

■ **Example 1.6** Let $X_i \ldots X_n \overset{iid}{\sim} N(\mu,\sigma^2)$ Find the information based on σ^2.
Write

$$\begin{aligned}\ell(\mu,\sigma^2) &= \log\left(\prod(2\pi\sigma^2)^{\frac{1}{2}}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}\right)\\ &= \frac{-n}{2}\log(2\pi\sigma^2) + \sum\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)\end{aligned}$$

If we try to use the expected information:
$I(\sigma^2) = E\left(\frac{\delta}{\delta\theta}\left(\frac{-n}{2}\log(2\pi\sigma^2) + \sum\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)\right)^2\right)$ which is a mess.
However, using the observed information:

$$\begin{aligned}I(\sigma^2) &= \frac{-n}{2\sigma^4} + \frac{1}{\sigma^6}E\left(\sum(X_i-\mu)^2\right)\\ &= \frac{-n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6}\\ &= \frac{n}{2\sigma^4}\end{aligned}$$

■

■ **Example 1.7** Continued from previous example.
Define $S^2 = \frac{1}{n-1}\sum(X_i-\bar{X})^2$
Note: $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$
Does S^2 acheive the C-R lower bound?

$$\begin{aligned}Var(S^2) &\geq \frac{\Psi'(\theta)}{I(\theta)}\\ \frac{2\sigma^4}{n-1} &\geq \frac{1}{I(\theta)}\\ &\geq \frac{2\sigma^4}{n}\end{aligned}$$

■

R Reread 7.3 and 6.2 in Casella and Berger

What would give equality? When W($\underline{X}$) is a linear function of $\frac{\delta}{\delta\theta}\log f_\theta(\underline{X})$ which leads to...

Corollary 1.3.4 — Attainment. Let $X_1,\ldots,X_n$ be iid f(x|θ), where f(x|θ) satisfies the conditions of the Cramer-Rao Theorem. Let $L(\theta|x) = \prod f(x_i|\theta)$ denote the likelihood function. If $W(X) = W(X_1,\ldots,X_n)$ is any unbiased estimator of $\tau(\theta)$, then W(X) attains the Cramer-Rao Lower Bound iff

$$a(\theta)\,(W(x) - \tau(\theta)) = \frac{\delta}{\delta\theta}\log L(\theta|x)$$

for some function a(θ).

Definition 1.3.5 — Efficient. Any estimator that achievies the Cramer-Rao bound is called an **efficient estimator**.

■ **Example 1.8** $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^s)$ where both are unknown. Let $\hat{\mu}_{ML} = \bar{X}$ and $\bar{\sigma^2}_{ML} = \frac{1}{n}\sum(x_i - \bar{X})^2$. Are these estimators efficient?

$$\begin{aligned}
Var(\mu_{ML}) &\geq \frac{1}{I(\mu)} \\
&= \frac{\sigma^2}{n} \\
I(\mu) &= nI_1(\mu) \\
&= nE\left[\left(\frac{\delta}{\delta\mu}\log f_\theta(x_i)\right)^2\right] \\
&= nE\left[\left(\frac{\delta}{\delta\mu}\frac{-1}{2}\log\left(2\pi\sigma^2\right) - \frac{(x_1-\mu)^2}{2\sigma^2}\right)^2\right] \\
&= nE\left[\left(\frac{(x_1-\mu)}{\sigma^2}\right)^2\right] \\
&= \frac{n}{\sigma^4}E(x_1^2 - 2x_1\mu + \mu^2) \\
&= \frac{n}{\sigma^4}(\sigma^2 + \mu^2 - 2\mu^2 + \mu^2) \\
&= \frac{n}{\sigma^2}
\end{aligned}$$

In this case, $\text{Var}(\mu_{ML}) = \frac{1}{I(\mu)}$, thus the sample mean is an efficient estimator.

For σ^2,

$$
\begin{aligned}
\mathrm{Var}(\hat{\sigma^2}_{ML}) &= \mathrm{Var}(\frac{1}{n}\sum(x_i-\bar{X})^2) \\
&= \frac{2(n-1)\sigma^2}{n^2} \\
I_1(\theta=\sigma^2) &= E\left[\left(\frac{\delta}{\delta\theta}\frac{-1}{2}\log(2\pi\theta)-\frac{(x_1-\mu)^2}{2\theta}\right)^2\right] \\
&= E\left[\left(\frac{-1}{2\theta}+\frac{(x_1-\mu)^2}{2\theta^2}\right)^2\right] \\
&= \mathrm{Var}\left[\frac{-1}{2\theta}+\frac{(x_1-\mu)^2}{2\theta^2}\right] \\
&= \mathrm{Var}\left[\frac{(x_1-\mu)^2}{2\theta^2}\right] \\
&= \frac{1}{4\theta^4}\mathrm{Var}\left[(x_1-\mu)^2\right] \\
&= \frac{1}{4\theta^4}\left[E(x_1-\mu)^4-\theta^2\right] \\
&= \frac{3\theta^2-\theta^2}{4\theta^4} \\
&= \frac{1}{2\sigma^4} \\
\left[\Psi'(\theta)\right]^2 &= \left[E\left(\frac{1}{n}\sum(x_i-\bar{x})\right)\right]^2 \\
&= \frac{(n-1)\theta}{n}
\end{aligned}
$$

So this estimator does not achieve the lower bound and thus is not efficient.

■

Issues with CR Bound:

(i) The C-R bound need not be attained by our favorite estimators.

(ii) Our assumptions do not always hold, and hense can not be used to apply the theorem. Especially when the range depends on a parameter.

For an alternative approach to finding the UMVUE, see the Rao-Blackwell and Lehmann -Sheffe Theorems.

1.3.3 Loss Function Optimality

Definition 1.3.6 — Loss. L(θ, W($\underline{x}$)) assigns a nonnegative real value called the **loss** to our decision to estimate θ by W($\underline{X}$). General context: Decision Theory.

Typically L(θ,θ) = 0 because nothing is lost if your decision is exactly correct.

■ **Example 1.9**

$$
\begin{aligned}
L(\theta, W(\underline{X})) &= (\theta - W(X)^2) \qquad \text{square error loss} \\
&= |\theta - W(X)| \qquad \text{absolute error loss} \\
&= \frac{W(X)}{\theta} - 1 - \log\frac{W(X)}{\theta} \qquad \text{Stein's loss}
\end{aligned}
$$

■

Definition 1.3.7 — Risk. Average loss for a given estimator. **Risk** of estimating θ by W($\underline{X}$) is

$$R(\theta, W) = E\left(\theta, W(\underline{X})\right)$$

Goal: We want to construct δ such that risk is minimized for every $\theta \in \Theta$.
Two Ways:

1. Restrict the space of estimators.
2. Allow all possible estimators but use a weaker notation of optimality
   (a) Bayesian approach
   (b) MiniMax approach

Definition 1.3.8 — Bayes Risk. Recall we previously define risk as a function of a parameter and estimator - E(Loss). Now that θ can be random, we may define **Bayes risk** as the average risk.

$$r_\pi(\delta) = \int R(\theta, \delta)\pi(\theta)d\theta = E(R(\theta, \delta)) = E(L(\theta, \delta(\underline{X})))$$

Definition 1.3.9 — Bayes Estimator. For a give loss function, we define the **Bayes estimator** as the minimizer of the Rayes Risk.

Note: if δ^* minimizes $E(L(\theta, \delta(\underline{X}))|\underline{X} = x)$ for (almost) all x then δ^* is the Bayes Estimator.
We can conclude that under square error loss, the posterior mean is the Bayes Estimator.
Wrap up: How to choose prior? Even choosing family of distribution is hard. Considering restraints and conjugacy can help.

- Use data to estimate prio parameters
- Assign priors on our priors "hyper priors". Priorception.
- Use a Jeffery's prior: reparameterize

Exercise 1.2 If $X_1, \ldots, X_n$ are iid with mean μ and varience σ^2 what is

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

Note:

1. $\sum i = i^n\left(X_i - \bar{X})^2\right) = \sum(X_i^2) - n\bar{X}$
2. Var(X) = $E(X^2) - E(X)^2$
3. Var$(\frac{X_i}{n}) = \frac{\sigma^2}{n^2}$
4. Var$(\bar{X}) = \frac{\sigma^2}{n}$

Thus,

$$E(\sum X_i^2) = (\sigma^2 + \mu^2)n$$

$$-nE(\bar{X}^2) = -n(\mu^2 + \frac{\sigma^2}{n})$$

We may conclude, $\frac{1}{n-1}\sum(X_i - \bar{X})^2$ is an unbiased estimator of σ^2 called S^2. ■

Definition 1.3.10 — Minimax Risk. The minimum risk of an estimater δ:

$$\inf_{all\,\delta} \sup_{\theta \in \Theta} R(\delta, \theta)$$

The Minimax risk is the minimum of the max risks. We say δ^* is minimax if

$$\sup_{\theta\in\Theta} R(\delta^*,\theta) = \inf_{all\,\delta}\sup_{\theta\in\Theta} R(\delta,\theta)$$

Theorem 1.3.5 The Bayes risk, $r_\pi(\delta)$, satistifes

$$r_\pi(\delta) = \int R_\pi(\delta|\underline{x})m(\underline{x})d\underline{x}$$

where $m(\underline{x}) = \int f_\theta(\underline{x})\pi(\theta)d\theta$ and $R_\pi(\delta|\underline{x}) = \int l(\delta,\theta)\pi(\theta|\underline{x})d\theta$ is called the posterior risk. If $\delta^*(\underline{x})$ minimizes the risk, then it is a Bayes estimator

Proof.

$$\begin{aligned} r_\pi(\delta) &= \int_\Theta R(\theta,\delta)\pi(\theta)d\theta \\ &= \int_\Theta\int_{x^n} l(\theta,\delta)f_\theta(\underline{x})d\underline{x}\pi(\theta)d\theta \\ \pi(\theta|\underline{x}) &= \frac{f_\theta(\underline{x})\pi(\theta)}{\int f_\theta(\underline{x})\pi(\theta)d\theta} \\ &= \frac{f_\theta(\underline{x})\pi(\theta)}{m(x)} \end{aligned}$$

■

Theorem 1.3.6 Suppose:

a) $L(\theta,\delta) = (\theta-\delta)^2$. Then the Bayes estimator is the posterior mean.

b) $L(\theta,\delta) = |\theta-\delta|$. Then the Bayes estimator is the posterior median.

c) $L(\theta,\delta) = \begin{cases} 0, & |\theta-\delta| \le \varepsilon \\ 1, & |\theta-\delta|\ \varepsilon \end{cases}$, then the Bayes estimator is the midpoint of an interval of length 2ε where $T(\theta|\underline{x})$ is maximized.

d) $L(\theta,\delta) = \begin{cases} 0, & \theta = \delta| \\ 1, & \theta \neq \delta \end{cases}$, then the Bayes estimator is the maximum aposterior.

1.3.4 Methods of Generating Reasonable Estimators

Epmirical Bayes - parameters of the prior are estimated from data.
Hierarchichal Bayes - we impose a hierarch of priors on the parameters of θ.
Choice of Prior: We want the prior impose **non-informative prior** where each value of the parameter is equiprobably. However, these priors could be **improper**, i.e. they don't integrate to 1 (in fact, not integrable at all).

Definition 1.3.11 — Jeffreys Prior. $\pi(\theta) \propto \sqrt{I(\theta)}$ If $\int\sqrt{I(\theta)}d\theta < \infty$ then π is a proper prior on θ. Else, we treat it as an improper prior, which is often the case for Jeffreys priors.

R

1. Jeffreys priors works well for single parameter models ,but has some issues while dealing with multiple parameters.
2. In general, Jeffreys prior is not a conjugate prior, but can be obtained as a limit of a conjugate prior family.

Definition 1.3.12 — Least Favorable Prior. A prior, π, is **least favorable** if

$$r_\pi(\delta^*) \geq r_\pi(\delta^*)$$

for all priors, π'

Theorem 1.3.7 Suppose π is a prior on Θ such that

$$\begin{aligned} r_\pi(\delta^*) &= \int R(\theta, \delta^*)\pi(\theta)d\theta \\ &= \sup_\theta R(\theta, \delta^*)(*) \end{aligned}$$

Then,

a) δ^* is minimax
b) π is the lease favorable prior.
c) If δ^* is unique Bayes Estimator, then it is also unique minimax estimator.

Question: When is (*) satisfied?

1. $R(\theta, \delta^*)$ is constant with respect to θ (i.e. does not depend on θ).
2. $\Omega = R(\theta, \delta^*) = \sup_{\theta^p rime} R(\theta', \delta^*)$. If $\pi(\Omega) = 1$ then (*) is satisfied.

■ **Example 1.10** $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 is known. Is $\bar{X}$ minimax under squared error loss?

Let $\pi(\mu) \sim N(\tau, b^2)$

δ^* is the Bayes estimator given by the posterior mean.

$$\delta^*(\underline{x}) = \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\tau}{b^2}}{\frac{n}{\sigma^2} + \frac{1}{b^2}}$$

As $b \to \infty$, $\delta^*(\underline{x} \to \bar{X}$. ($\pi$ is an improper prior for μ implies it's the Jeffreys prior for μ).

$$
\begin{aligned}
R(\mu,\delta^*) &= E[L(\mu,\delta^*)] \\
&= E[(\mu-\delta^*)^2] \\
&= E[(\alpha\bar{x}+(1-\alpha)\tau-\mu)^2] \\
&= E[\alpha^2(\bar{x}-\mu)^2+(1-\alpha)^2(\tau-\mu)^2+2\alpha(\bar{x}-\mu)(1-\alpha)(\tau-\mu)] \\
&= \alpha^2 * E(\bar{x}-\mu)^2+(1-\alpha)^2(\tau-\mu)^2 \\
&= \alpha^2\mathrm{Var}(\bar{x})+(1-\alpha)^2(\tau-\mu)^2 \\
&= \frac{\alpha^2\sigma^2}{n}+(1-\alpha)^2(\tau-\mu)^2 \\
r_\pi(\delta^*) &= \int R(\mu,\delta^*)\pi(\mu)d\mu \text{(Note: Risk from line above.)} \\
&= \int [\frac{\alpha^2\sigma^2}{n}+(1-\alpha)^2(\tau-\mu)^2]\pi(\mu)d\mu \\
&= \frac{\alpha^2\sigma^2}{n}+(1-\alpha)^2b^2
\end{aligned}
$$

Substituting for α, we get

$$
\begin{aligned}
r_\pi(\delta^*) &= \frac{1}{\frac{n}{\sigma^2}+\frac{1}{b^2}} \\
&= \int R(\mu,\delta)\pi(\mu)d\mu \\
&\leq \sup_\mu R(\mu,\delta)
\end{aligned}
$$

In other words,

$$
\begin{aligned}
\sup_\mu R(\mu,\delta) &\geq \frac{1}{\frac{n}{\sigma^2}+\frac{1}{b^2}} \\
&\geq \sup_b \left(\frac{1}{\frac{n}{\sigma^2}+\frac{1}{b^2}}\right) \\
&= \frac{\sigma^2}{n}, \forall\delta \\
&\implies \inf_\delta \sup \mu R(\mu,\delta) = \frac{\sigma^2}{n} = R(\mu,\bar{x}) = \sup_\mu R(\mu,\bar{x})
\end{aligned}
$$

Therefore, $\bar{x}$ is minimax. ∎

Definition 1.3.13 — Admissibility. An estimator δ is said to be inadmissable if there exists δ' that dominates it, i.e. $\mathrm{R}(\theta,\delta) \geq R(\theta,\delta')$ for all θ and $R(\theta,\delta) > R(\theta,\delta')$ for some θ. If no such δ' exists, then δ is **admissible**.

a) UMVUE need not be admissible as biased estimators can yield smaller MSE.
b) But the situation is different when dealing with minimax estimators.
c) If δ is a minimax estimator and it is dominated by δ', then δ' is also minimax and therefore we prefer δ'.
d) So it is important to ascertain whether a given minimax estimator is admissable or not.

Theorem 1.3.8 Any unique Bayes estimator is **admissible**.

$$\inf_{\delta} r_\pi(\delta)$$

Proposition: If an estimator is unique minimax, then it is admissible.
Proposition: If δ has a constant risk and is admissible then it is minimax.

1.4 Practice Problems

Chapter 7 Exercises (pp. 355 to 367): 7.2, 7.6, 7.7, 7.10, 7.12, 7.13, 7.14, 7.19, 7.20, 7.21, 7.22, 7.23, 7.25, 7.30, 7.37, 7.38, 7.44, 7.47, 7.48, 7.49, 7.52

Problem 1.1 (C&B 7.2) Let $X_1,\ldots,X_n$ be a random sample from a Gamma(α,β) population.
(a) Find the MLE of β, assuming α is known.
(b) If α and β are both unknown, there is no explicit formula for the MLEs of α and β, but the maximum can be found numberically. The result in part (a) can be used to reduce the problem to the maximization of a univariate function. Find the MLEs for $\alpha$and β for the data in Exercise 7.10(c).

Solution.
(a) Gamma distribution pdf: $\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$

$L(\beta) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}[\prod x]^{\alpha-1}e^{-\beta\sum x}$

$\ell(\beta) = n\alpha\log\beta - n\log\Gamma(\alpha) + (\alpha-1)\log(\prod x_i) - \beta\sum x_i$

$\frac{\delta\ell}{\delta\beta} = \frac{n\alpha}{\beta} - \sum x_i$

$\hat{\beta} = \frac{n\alpha}{\sum x_i}$

R The density of Gamma can also be in the form with $\frac{1}{\beta}$ which would mean $\hat{\beta}$ would be $\frac{\sum x_i}{n\alpha}$.

(b) The new likelihood with $\hat{\beta}$ plugged in is:

$$L(\beta) = \frac{(\frac{n\alpha}{\sum x_i})^{n\alpha}}{\Gamma(\alpha)^n}[\prod x]^{\alpha-1}e^{-(\frac{n\alpha}{\sum x_i})\sum x}$$

To solve...use a computer.

Problem 1.2 (C&B 7.6) Let $X_1,\ldots,X_n$ be a random sample from the pdf

$$f(x|\theta) = \theta x^{-2}, 0 < \theta \le x < \infty$$

(a) What is a sufficient statistic for θ?
(b) Find the MLE of θ.
(c) Find the method of moments estimator of θ?

Solution.
(a) Use Factorization Theorem.
$L(\theta) = \theta^n \prod(x_i^2) \prod I\{\theta \leq x < \infty\}$ where $\prod I\{\theta \leq x < \infty = x_{(1)}$
Thus the sufficient statistic is the minimum x.

(b) L(θ|x) = $\theta^n \prod(x_i^2) \prod I\{\theta \leq x < \infty\}$. θ^n is increasing in θ. The second term does not involve θ. So to maximize $L(\theta|x)$, we want to make θ as large as possible. But because of the indicator function, $L(\theta|x) = 0 if \theta > x(1)$. Thus, $\hat{\theta} = x_{(1)}$.

(c) $E(X) = \int_\theta^\infty \theta x^{-1} dx = \theta \log x|_\theta^\infty = \infty$. Thus the method of moments estimator of θ does not exist.

Problem 1.3 Let $X_1, \ldots, X_n$ be iid with one of two pdfs. If θ = 0, then

$$f(x|\theta) = \begin{cases} 1, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$f(x|\theta) = \begin{cases} \frac{1}{2\sqrt{x}}, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Find the MLE of θ.

Solution.

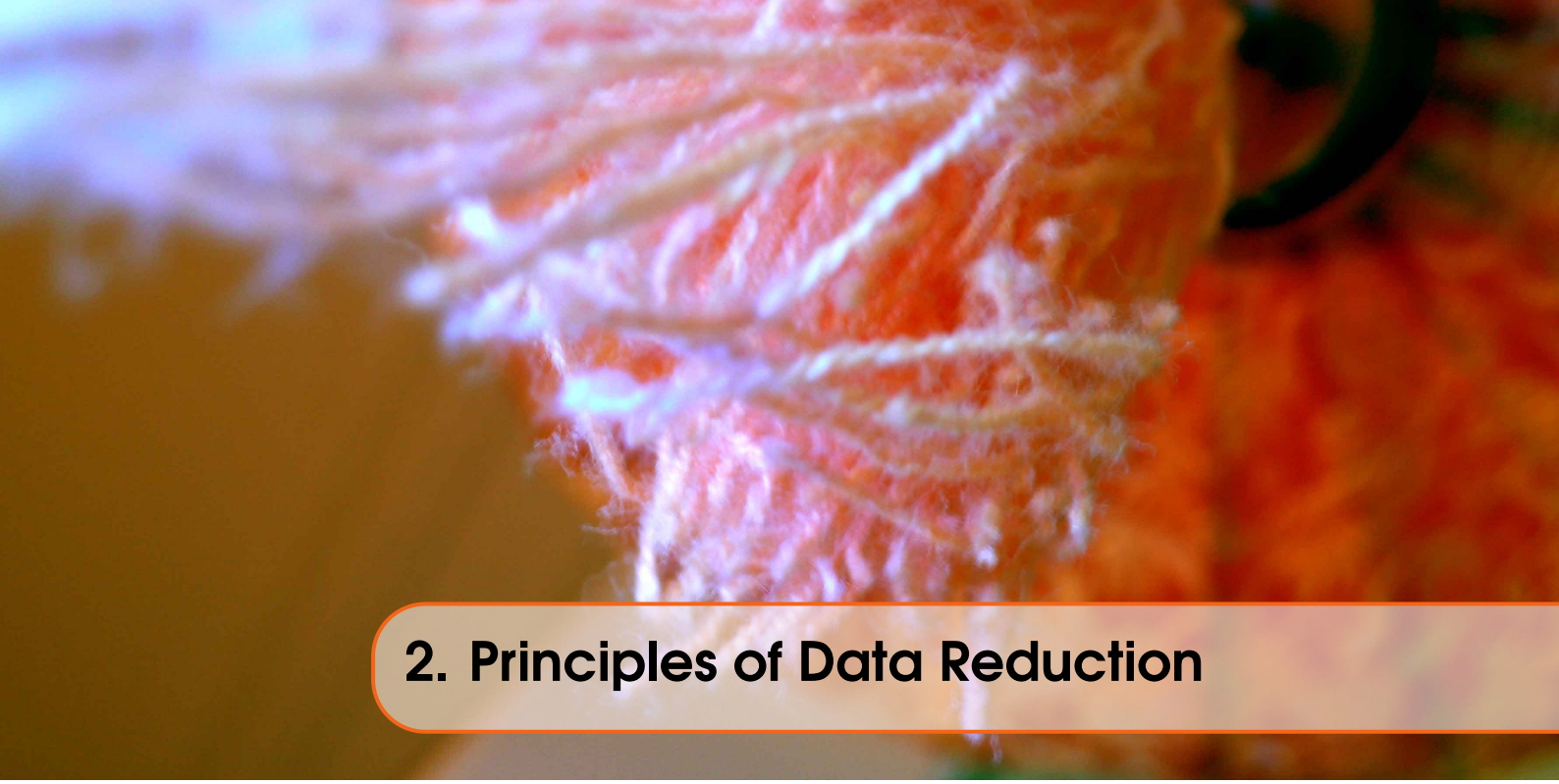
$$L(0|x) = \begin{cases} 1, & \text{if } 0 < x_i < 1 \\ 0, & \text{otherwise} \end{cases}$$

while if $\theta = 1$, then

$$L(1|x) = \begin{cases} \prod \frac{1}{2\sqrt{x}}, & \text{if } 0 < x_i < 1 \\ 0, & \text{otherwise} \end{cases}$$

Thus,

$$\hat{\theta}_{MLE} = \begin{cases} 1, & \text{if } \prod \frac{1}{2\sqrt{x}} \\ 0, & \text{otherwise} \end{cases}$$



2. Principles of Data Reduction

2.1 Sufficiency Principle

Definition 2.1.1 — Sufficient. T(X) is **sufficient** for θ if the distribution of X|T($\underline{X}$) does not depend on θ. Note that sufficient statistics are not unique.

Theorem 2.1.1 — Factorization Theorem. If we have $\underline{X} \sim f_\theta(\underline{x})$ then T is sufficient if and only iff we can write f_θ as

$$g(T(x),\theta)h(x)$$

for some g and h.

■ **Example 2.1** $X_1,\ldots,X_n \overset{iid}{\sim} \text{Bern}(\theta)$
pdf: $\theta^x(1-\theta)^{1-x}$
Joint pdf = $f_\theta(x) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$
So, T($\underline{x}$) = $\sum x_i$ ■

Q: What is the connection to Section 7.3?
A:

$$\begin{aligned} Var_\theta(W(\underline{X})) &= Var\left[E(W(X)|T(X))\right] + E\left[Var(W(X)|T(X))\right] \\ &\geq Var\left[E(W(X)|T(X))\right] \end{aligned}$$

Note: $E_\theta\left\{E_\theta[W(X)|T(X)]\right\} = E_\theta(W(X))$
$E_\theta[W(X)|T(X)]$ does not depend on θ. So this is a legitimate estimator since T(X) is sufficient and its varience is smaller than any estimator with same mean.

General Idea of Sufficiency: If T(X) is sufficient for θ, then all information in X about θ is captured in T(X).

More technically, conditioning T(X), the remaining randomness does not depend on θ.

R Know Thm 6.2.6 - Factorization Theorem

Definition 2.1.2 — Minimal Sufficient. T(X) is **minimal sufficient** if for any other sufficient statistics, $T'(X)$, $T(X)$ is a function of $T'(X)$. Note: "T(X) is a function S(X)" means that S(x) = S(y) implies T(x) = T(y).

■ **Example 2.2** $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$. Find possible $T(\mu,\sigma^2)$.

$$\begin{aligned} f(\underline{x}) &= \prod (2\pi\sigma^2)^{\frac{1}{2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{\frac{-n}{2}} \exp\left\{\frac{-\sum x_i^2 + 2\mu\sum x_i - n\mu^2}{2\sigma^2}\right\} \end{aligned}$$

Therefore our pair of sufficient statistics are $\sum x_i^2$ and $\sum x_i$. But are the minimal? ■

Theorem 2.1.2 $T(\underline{x}) = T(\underline{y})$ if and only if $\frac{f_\theta(\underline{x})}{f_\theta(\underline{y})}$ does not depend on θ, implying T($\underline{x}$) is minimal sufficient statistic.

R Minimal sufficient statistics are not unique (any 1-1 transformation of a minimal sufficient statistics is also minimal sufficient). A sufficient statistic may have components that do not contain any information about θ. Such components are called **ancillary statistics**.

2.2 Exponential Families

Definition 2.2.1 Exponential Family An **exponential family** of densities is the set of densities given by

$$f_{\underline{\theta}}(\underline{x}) = h(\underline{x})c(\underline{\theta})\exp\left\{\sum w_i(\underline{\theta})T_i(\underline{x})\right\}$$

As θ takes values in some set Θ, this equation may be rewritten as

$$f_{\underline{\eta}}(\underline{x}) = \exp\left\{\underline{\eta}^T \underline{T}(\underline{x}) - A(\underline{\eta})\right\} h(\underline{x}) I_B(\underline{x})$$

The latter equation is called the **canonical/natural parameterization**.

R The $T_i/T(x)$ terms are sufficient statistics.

■ **Example 2.3** Deriving the canonical parameterization for a Normal distribution.

$$\begin{aligned} f_{(\mu,\sigma^2)} &= k\frac{1}{\sqrt{2\sigma^2}}\exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\} \\ &= k\exp\left\{\frac{1}{2\sigma^2}(-x^2) + \frac{\mu}{\sigma^2}(x) - \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log((2\sigma^2))\right\} \\ &= k\exp\left\{\eta_1(-x^2) + \eta_2(x) - \frac{\eta_2^2}{4\eta_1} - \frac{1}{2}\log\eta_1\right\} \end{aligned}$$

$\eta_1 = \frac{1}{2\sigma^2} \Leftrightarrow \sigma^2 = \frac{1}{2\eta_1}$
$\eta_2 = \frac{\mu}{2\sigma^2} \Leftrightarrow \mu = \frac{\eta_2}{2\eta_1}$
$A(\underline{\eta}) = \frac{\eta_2^2}{4\eta_1} - \frac{1}{2}\log\eta_1$

■

Cool fact: The partial derivatives of A(η) gives the expectations of T. The second partials give the covarience of T.
Observe $T(x) = (-x^2, x)$

$$\begin{aligned} E(-x^2) &= \frac{\delta}{\delta\eta_1} A(\underline{\eta}) \\ &= \frac{-\eta_2^2}{4\eta_1^2} - \frac{1}{2\eta_1} \\ &= \frac{-\mu^2 * 4\sigma^4}{4\sigma^4} - \frac{2\sigma^2}{2} \\ &= -\mu^2 - \sigma^2 \end{aligned}$$

$$\begin{aligned} E(x) &= \frac{\delta}{\delta\eta_2} A(\underline{\eta}) \\ &= \frac{\eta_2}{2\eta_1} \\ &= \mu \end{aligned}$$

$$\begin{aligned} Var(x^2) &= \frac{\delta^2}{\delta^2\eta_1} \\ &= \frac{\eta_2^2}{2\eta_1^3} + \frac{1}{2\eta_2} \\ &= 4\mu^2\sigma^2 + 2\sigma^4 \end{aligned}$$

2.2.1 Distribution Families

Poisson

$$\begin{aligned} f_\lambda &= e^{-\lambda}\frac{\lambda^x}{x!} I\{x \in \mathbb{N}\} \\ &= \exp\{(\log\lambda)x - \lambda\} \\ \eta &= \log\lambda \\ T(\underline{x}) &= x \\ A(\eta) &= \lambda = e^{\log\lambda} = e^\eta \end{aligned}$$

Beta

$$\begin{aligned} f_{\alpha,\beta}(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I\{0 < 1x < 1\} \\ &= \exp\{\log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) + (\alpha-1)\log x + (\beta-1)\log(1-x) I\{0 < x < 1\}\} \\ \eta_1 &= \alpha - 1 \\ \eta_2 &= \beta - 1 \\ T_1(x) &= \log(x) \\ T_2(x) &= \log(1-x) - x\beta + \alpha\log(\beta) - \log(\Gamma(\alpha)) \end{aligned}$$

Gamma

$$
\begin{aligned}
f_{\alpha,\beta}(x) &= \frac{\beta^{\alpha}x^{\alpha-1}}{\Gamma(\alpha)}\exp\{-x\beta\}I\{x>0\} \\
&= \exp\{(\alpha-1)\log(x)-\beta(x)+(\eta_1+1)\log(-\eta_2)+\log(\Gamma(\eta_1+1))\} \\
\eta_1 &= \alpha-1 \\
\eta_2 &= \beta \\
T_1(x) &= \log(x) \\
T_2(x) &= x \\
A(\eta) &= (\eta_1+1)\log(-\eta_2)+\log(\Gamma(\eta_1+1))
\end{aligned}
$$

Binomial

$$
\begin{aligned}
f_p(x) &= \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}I\{0\ldots,n\} \\
&= \binom{n}{x}\exp\left\{x(\log\frac{p}{1-p})-(-n\log(1-p))\right\} \\
\eta &= \log\frac{p}{1-p} \\
T(x) &= x \\
A(\eta) &= -n\log(1-p)
\end{aligned}
$$

Why is E(T(x)) = $\frac{\delta}{\delta\eta}A(\eta)$?
Notice: $e^{A(\eta)}$ must be the integral of $\exp\{\underline{\eta}^T\underline{T(x)}-A(\underline{\eta})\}h(\underline{x})$.
Differentiating on both sides and using the fact that for an expoonential family we may switch integral and derivative, we obtain:

$$
\begin{aligned}
\frac{\delta}{\delta\eta}A(\eta)e^{A(\eta)} &= e^{A(\eta)}\int\exp\{\eta^T T(x)\}h(x)[T(x)]e^{-A(\eta)} && (2.1)\\
&= E(T(x)) && (2.2)\\
& && (2.3)
\end{aligned}
$$

Recall, $f_\eta(x)$ must always integrate to 1!

R Read Casella & Berger Chapter 7.2

2.3 Complete Statistic

Recall, minimal statistics contain no redundant or unnessicary information. What might redundant information look like? From T(x) we may construct g_1 and g_2 such that $E_\theta(g_1(T)) = E_\theta(g_2(T))$ for all θ which we want to avoid.

Definition 2.3.1 — Complete Statistic. (T) is obtained only if $g=0$ gives $E_\theta(g(T))=0$ for all θ.

Theorem 2.3.1 Complete and sufficient inplies minimal statistics.

Theorem 2.3.2 If $f_\theta(\underline{x})\exp\left\{\eta^T T(x) = A(\eta)\right\}h(x)$ is a cononical exponential family of full rank then T is a complete and sufficient statistic.

Definition 2.3.2 Full Rank **Full rank** means
Parameter space H contains an open set
- T does not satisfy any linear constraint (i.e. the T are linearly independent).

Definition 2.3.3 — Open Set. Let k be dimention of η. To say that H contains an **open set** means there exists $\begin{cases}\varepsilon > 0\\ \eta_0 \in H\end{cases}$ such that $B(\eta_0;\varepsilon) \subset H$

Theorem 2.3.3 — Basu's Theorem. Any complete and sufficient statistic (minimum sufficient) is independent of any ancillary statistic (that is, any statistic whose distribution does not depend on θ).

2.3.1 Sufficiency and Unbiasness

Suppose $E_\theta(W(X)) = \theta$ for all θ, thus W(X) is unbiased. Then Var$[E(W(X)|T)] \leq Var(W(X))$.
But what if T is not sufficient? Wel...then you won't get an estimator.

■ **Example 2.4** Suppose that $X \sim Unif(0,\theta)$. Find the UMVUE of $\sin\theta$.

Solution.
pdf: $f(x) = \frac{1}{b-a} = \frac{1}{\theta}$
E(X)= $\frac{\theta}{2}$
Thus 2X is unbiased for θ. But is $\sin(2X)$ unbiased for $\sin(\theta)$?
NO!

What about cos(2X)?

$$\begin{aligned}E(\cos(2X)) &= \int_1^\theta \cos(2X)\frac{1}{\theta}dx\\ &= \frac{1}{\theta}\left(\frac{\sin(2X)}{2}\right)|_0^\theta\\ &= \frac{\sin(2\theta)}{2\theta}\end{aligned}$$

No! Blarg. Let's bust out some calculus.

$$\begin{aligned}\int_0^\theta \textbf{estimator(x)}dx &= \theta\sin(\theta)\\ \textbf{estimator} &= \sin\theta + \theta\cos(\theta)\\ &= x\sin(x) + x\cos(x)\end{aligned}$$

■

Theorem 2.3.4 — Rao-Blackwell. Let W(X) be an extimator with $E_\theta(W(x)) = \tau(\theta)$. If T(X) is a sufficient statistic, then an alternative, unbiased estimator of $\tau(\theta)$ whose varience is

uniformly not worse than that of W(X) is

$$\phi(X) = E_\theta(W(X)|T(X))$$

.

Proof. Based on conditional arguments.

$$E(W) = E(E(W(X)|T(X)))$$

Var(W)=Var(Cond Exp)+ E(Cond Var) ■

Theorem 2.3.5 If W(X) is UMVUE for $\tau(\theta)$ then it is the unquie UMVUE, aka the best unbiased estimator for all θ.

Theorem 2.3.6 7.3.20 If $E_\theta(W(X)) = \tau(\theta)$, then W(X) is UMVUE for $\tau(\theta)$ iff W(X) is uncorrelated with any S(X) such that $E_\theta(S(X)) = 0$.

Theorem 2.3.7 — Lehmann-Scheffe Theorem. If T(X) is a complete and sufficient statistic, then $\phi(T)$ is the unique UMVUE of its expectation (see Rao-Blackwell).

Definition 2.3.4 A funciton $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex if domain (ϕ) is a convex set and if for all $x, y \in dom(\phi)$ and $0 \leq \alpha \leq 1, \phi(dx + (1-\alpha)u) \leq \alpha\phi(x) + (1-\alpha)\phi(y)$. If strict equality holds for all $x \neq y$, then ϕ is strictly convex.

Characterization of Convexity:

1. First-order conditions: suppose ϕ is differentiable (i.e. the gradient $\nabla\phi$ exists at each point in dom(ϕ) which is open). Then ϕ is convex iff dom(ϕ) is convex and $\phi(y) \geq \phi(x) + \nabla\phi^T(x)(y-x), \forall x, y$.
2. Second-order conditions: suppose f is twice differentiable. That is, the Hessian $\nabla^2\phi$ exists at each point in dom(ϕ), which is open. Then ϕ is convex iff dom(ϕ) is convex and $\nabla^2\phi$ is positive semidefinite.

Definition 2.3.5 — Jenson's Inequality. If h(x) is a convex function then

$$h(E(x)) \leq E(h(x))$$

for any random variable x.

Theorem 2.3.8 If ϕ is convex, then $\phi(E(X)) \leq E(\phi(X))$ and the strict inequality holds if ϕ is strictly convex.

Corollary 2.3.9 Let X be a idscrete random variable with $P(X = x_i) = \alpha_i, i-1, ...l$. Apply the above theorem.

Uniqueness of δ^*:

$$\delta^* \in \arg\inf_\delta r_\pi(\delta)$$

We are done if $r_\pi(\delta)$ is strictly convex for $0 < \alpha < 1$, consider

$$\begin{aligned} r_\pi(\alpha\delta + (1-\alpha)\delta_2) &= \int R(\theta, \alpha\delta_1 + (1-\alpha)\delta_2)\pi(\theta)d\theta \\ &= \int E\left[L(\theta, \alpha\delta(1 + (1-\alpha)\delta_2))\right]\pi(\theta)d\theta \\ &< E\left[\alpha L(\theta,\delta) + (1-\alpha)L(\theta,\delta_2)\right]\pi(\theta)d\theta \\ &= \int (\alpha E(L(\theta,\delta_1) + (1-\alpha)E(L(\theta,\delta_2)))) \\ &= \alpha r_\pi(\delta_1) + (1-\alpha)r_p i(\delta_2) \end{aligned}$$

Thus, $r_\pi$is strictly convex, which implies that d^* is the unique minimizer.
Generalized Version of Rao-Blackwell Theorem Let w be an estimator of $g(\theta)$ and T be a sufficient statistic. Let $L(\theta,\delta)$ be a strictly convex function in its second argument.
Suppose $E(W) < \infty$ and $R(\theta,W) = E(L(\theta,W)) < \infty$. Define $\eta(T) = E(W|T)$. Then $R(\theta,\eta(T)) < R(\theta,W)$ unless $\eta(T) = W$ with probability 1.

Theorem 2.3.10 Suppose T is complete and sufficient. Then for every $g(\theta)$ that can be estimated by an unbiased estimator, there is one and only one unbiased estimator that is a function of T. In addition, this estimator uniformly minimizers the risk for any loss function $L(\theta,\delta)$ that is stictly convex in δ. In particular, this estimator is UMVUE.

We say $\hat{\theta}$ is **efficient** if it achieves equality in the information bound. We say $\hat{\theta}_{ML}$ is **asymptotically efficient** if as $n \to \infty$, $\text{Var}(\hat{\theta}_{ML}) \to$ the CR bound.

2.4 Major/Minorization Mini/Maximization

Every EM Algorhythm is special case of a MiMax.
EM = Expectation Maximization

R Read 7.2.4 in Casella & Berger

2.4.1 Bharath's Notes

Minoroization - Maximization

1. Constructing a minorization function.

$$f(x) \geq g(x,y), \forall x,y \in \Omega$$
$$f(x) = g(x,x), \forall x \in \Omega$$

2. $X^{(t+1)} \in \text{argmax}_{x\in\Omega\subset\mathbb{R}^d} g(x,x^{(t)})$
3. Return to step 1.

Claim: $f(x^{t+1}) \geq f(x^{(t)})$

Proof.

$$\begin{aligned} f(x^{t+1}) &\geq g(x^{(t+1)}, x^{(t)}) \\ &\geq g(x^{(t)}, x^{(t)}) \\ = f(x^{(t)}) \end{aligned}$$

■

EM Algorithm: Goal: Maximize $L(\theta) = f_\theta(x_1,\ldots,x_n)$ over $\theta \in \Theta \subset \mathbb{R}^d$.

Definition 2.4.1 The Kullbach-Liebler divergence between two densities p and q is defined as

$$\begin{aligned} KL(p||q) &= E_{x\sim q}(\log \frac{q(x)}{p(x)}) \\ &= \begin{cases} \int q(x)\log\frac{q(x)}{p(x)}dx, & if q << p \end{cases} \end{aligned}$$

2.4.2 Dave's Notes

■ **Example 2.5** $X_1,\ldots,X_n \overset{iid}{\sim} \text{Bern}(p), 0<p<1$

$$\begin{aligned} \hat{p}_{mm} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i}L(p) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ l(p) &= \sum x_i \log(p) + (n+\sum x_i)\log(1-p) \\ \frac{dl(p)}{dp} &= \frac{\sum x_i}{p} + \frac{n-\sum x_i}{(1-p)} \\ &\implies \hat{p}_{ML} = \frac{1}{n}\sum x_i \\ \frac{d^2 l(p)}{dp^2} &= \frac{-\sum x_i}{p^2} - \frac{n-\sum x_i}{(1-p)^2} < 0 \end{aligned}$$

and therefore $\hat{p}_{ML}$ maximizes L(p). ■

2.4.3 How EM Algorhythms Work

a) Choose a starting parameter, θ.
E-Step Construct the (minimizing) function.
b) Maximize this function of θ. The maximizer will be next value of θ_0. Call it θ_1.
c) Return to (b) as long as we haven't converged.

Given μ_0 we find after one EM iteration that

$$\begin{aligned} \mu_1 &= \frac{\sum Y_i}{n} + \mu_0(1-\frac{\mu}{n}) \\ &= \mu_0(1-\frac{\mu}{n}) + \frac{\sum Y_i}{\mu}(\frac{\mu}{n}) \end{aligned}$$

Thus, μ_1 always takes us $\frac{\mu}{n}$ of the way to the final answer at each iteration. This type of convergence is called **linear convergence**, which is a characteristic of EM algorhythems. The linear rate of convergence is governed by "amount of missingness" and is considered slow compared to other similar methods (e.g. Newton-Raphson is optimization method that enjoys quadradic convergence). However, EM algorhithem tend to trade more iterations for simpler iterations.

Suppose we wish to maximize a fuction that is a product of sums or integrals. Taking the log gives a sum of logs of sums/integrals:

$$\sum^{A} \log \sum^{B} s_{ab}(\theta)$$

To simplify, ignore the sumation over A and take $f(\theta) = \log \sum s_{ab}(\theta)$. Fix a θ_0. Define

$$W_{ob} \equiv \frac{S_b(\theta_0)}{\sum_{c_a}^{c} S(\theta_0)}$$

Goal: Maximize f(θ) $= \log \sum S_b(\theta)$ with fixed θ_0.
Claim: Define $Q_0(\theta) \equiv \sum^B W_{ob} * \log S_b(\theta)$.
Then $f(\theta) - f(\theta_0) \geq Q_0(\theta) - Q_0(\theta_0)$.

To verify this claim, notice that Jenson's inequality says

$$E(\log(\bullet)) \geq \log E(\bullet)$$

$$\begin{aligned} Q_0(\theta) - Q_b(\theta_0) &= \sum^B W_{ob} \log(\frac{S_b(\theta)}{S_b(\theta_0)}) \\ &\leq \log \sum^B W_{ob} \frac{S_b(\theta)}{S_b(\theta_0)} \\ &\leq \log \sum^B \frac{S_b(\theta_0)}{\sum S_c(\theta_0)} \frac{S_b(\theta)}{S_b(\theta_0)} \\ f(\theta) - f(\theta_0) &= \log \frac{S_b(\theta)}{S_c(\theta_0)} \end{aligned}$$

■ **Example 2.6** $X \sim f_\theta(x) = \lambda f_{\xi_1}(x) + (1-\lambda) f_{\xi_2}(x), \theta = (\lambda, \xi)$ Intuition: To generate X acoording to mixture density flip coin with Heads probability of λ. If H, generate $X \sim f_{\xi_1}(x)$. If T, generage $X \sim f_{\xi_2}$.
Write down the observed data likelihood, $\ell(\lambda, p_1, p_2)$. For a sample size, n, form a mixture of Binom(m, p_1) and Binom(m, p_2).

$$\sum \log \left(\lambda \binom{m}{\xi} p_1^{\xi_1} (1-p_1)^{m-\xi_1} + (1-\lambda) \binom{m}{\xi_2} p_2^{\xi_2} (1-p_2)^{m-\xi_2} \right)$$

Now impliment EM algorithm. ■

Suppose that "complete data" consists of x_{obs}&x_{miss}.
Using previous notation,

$$\begin{aligned} S_b(\theta) &= P_\theta \left(X_{obs} = x_{obs}, X_{miss} = b \right) \\ &\Rightarrow \sum^B S_b(\theta) = P_\theta \left(X_{obs} = xobs \right) \end{aligned}$$

$$\begin{aligned} W_{ob} &= \frac{P_{\theta_0} \left(X_{obs} = x_{obs}, X_{miss} = b \right)}{P_\theta \left(X_{obs} = x_{obs} \right)} \\ &= P_{\theta_0} \left(X_{miss} = b | X_{obs} = x_{obs} \right) \end{aligned}$$

$$Q_0(\theta) = E_{\theta_0} [\log P_\theta \left(X_{obs} = x_{obs}, X_{miss} = b \right) | X_{obs} = x_{obs}]$$

3. Hypothesis Testing

Definition 3.0.1 hypothesis testing We see that a hypothesis testing procedure is a rule that partitions the sampe space into we will accept (fail to reject) H_0 as true or not.

Let's consider a simple null hypothesis, i.e. a hypothesis of the form:

$$H_0 : \theta = \theta_0$$

We will consider three different ways to determine whether $\theta_{true} = \theta_0$

1. **Wald Test** $\hat{theta} - \theta_0$ - consider if this is large relative to the distribution. $\hat{\theta}$ should have if H_0 true. Recall, $\sqrt{n}(\hat{\theta} - \theta_0) \sim N(0, I^{-1}(\theta_0))$.
2. **Likelihood Ratio Test** Consider $\ell(\hat{\theta}) - \ell(\theta_0)$ and determine whether this is large relative to χ^2 distribution that it will approximately have under H_0.
3. **Rao Score Test** Consider $\nabla\ell(\theta_0)$ and determine if it is too far from 0 relative to it's true approximate distribution if H_0 true.

R Read 8.2.1 and 8.2.2

Theorem 3.0.1 If T(X) is sufficient for the family of distributions from which X is drawn and $\delta(X)$ is an unbiased estimater

Bibliography

Index