STAT 514 Lecture Notes

Dr. David Hunter

Copyright © 2014 John Smith

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at http://creativecommons.org/licenses/by-nc/3.0. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, March 2013

Contents

1 Point Estimation - Chapter 7 Casella & Berger 5

1.1 Introduction 5

1.2 Mean Squared Error 6

1.3 Best Unbias Estimator 7

1.4 Lost Function Optimality 9

1.5 Practice Problems 10

2 Principles of Data Reduction 13

2.1 Sufficiency Principle 13

3 Presenting Information 15

3.1 Table 15

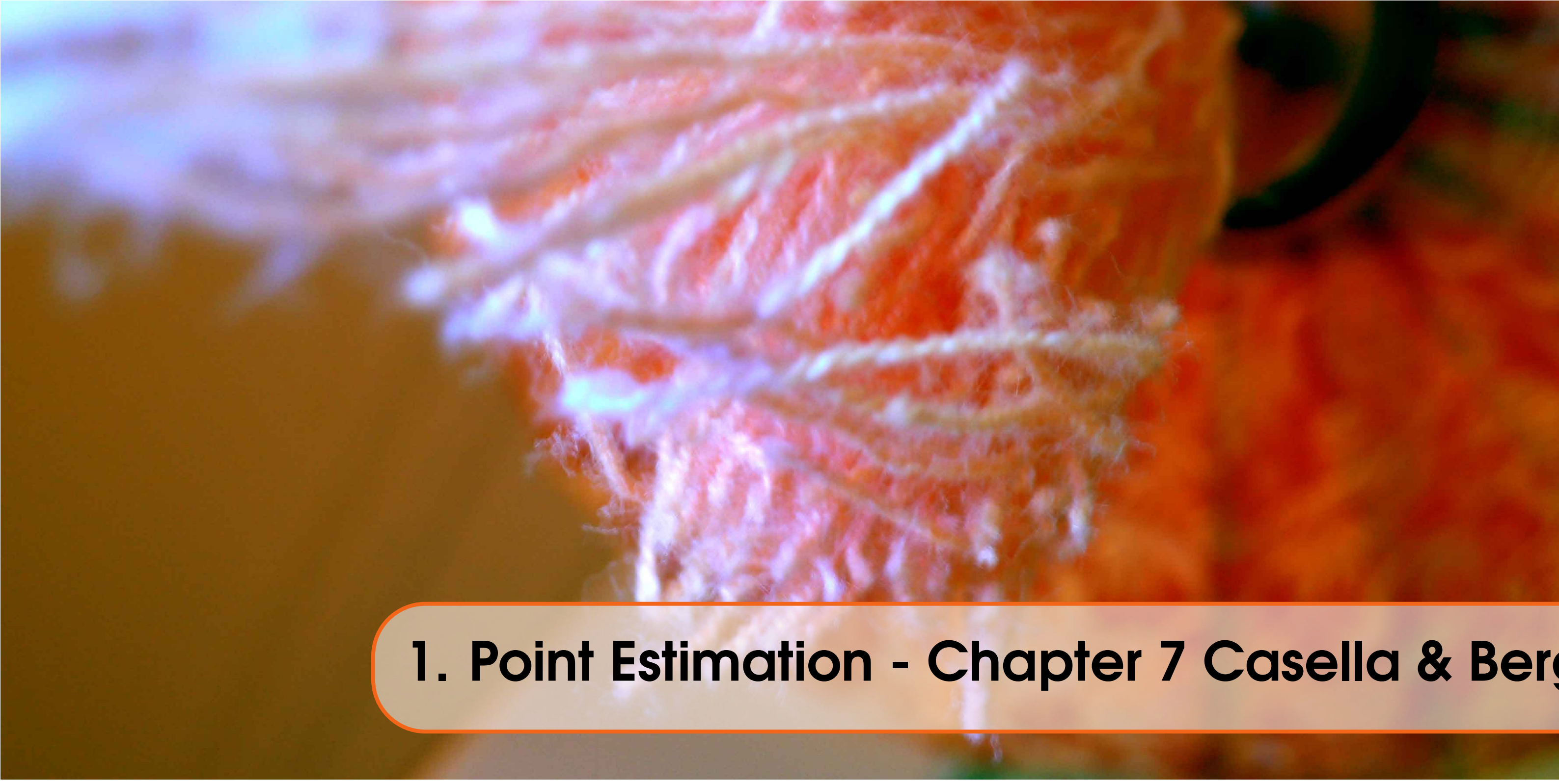
3.2 Figure 15

Bibliography 17

Books 17

Articles 17

Index 19



1. Point Estimation - Chapter 7 Casella & Berg

1.1 Introduction

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

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

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 $W(X_1, \dots, X_n)$ 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. $W(X_1, \dots, X_n)$ is the estimator, $W(x_1, \dots, x_n)$ is the fixed estimate.

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

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

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 Mean Squared Error

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

Definition 1.2.1 — 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 1.2.2 — 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 1.2** 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 1.2.1 $\text{MSE}(W) = \text{bias}^2 + \text{Var}(W)$

Proof:

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

1.3 Best Unbias Estimator

What does best mean? Answer: Minimum variance.

Recall: Given an esitmator $W(\underline{X})$ for θ,

$$MSE(\theta) = E(W(\underline{X}) - \theta)^2)$$

Definition 1.3.1 — 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

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

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)).

Theorem 1.3.1 — Cremer-Rao Inequality. Also: Information Inequality.

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

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

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

Assumptions:

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

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 1.3** 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 1.4** 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 1.3.2 — 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
>
> $$a(\theta)\left(W(x)-\tau(\theta)\right) = \frac{\delta}{\delta\theta}\log L(\theta|x)$$
>
> for some function a(θ).

1.4 Lost Function Optimality

Definition 1.4.1 — 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 1.5**

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

■

Definition 1.4.2 — Risk. **Risk** of estimating θ by W($\underline{X}$) is

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

Exercise 1.2 If $X_1,\dots,X_n$ 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. ■

Theorem 1.4.1 — Cauchy-Scwarz Inequality.

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

In terms of E(X) if A & B have μ = 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 $D^2\geq 0$.

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

1.5 Practice Problems

Problem 1.1 (C&B 7.2) Let $X_1,\dots,X_n$ 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 $\frac{\sum x_i}{n\alpha}$.

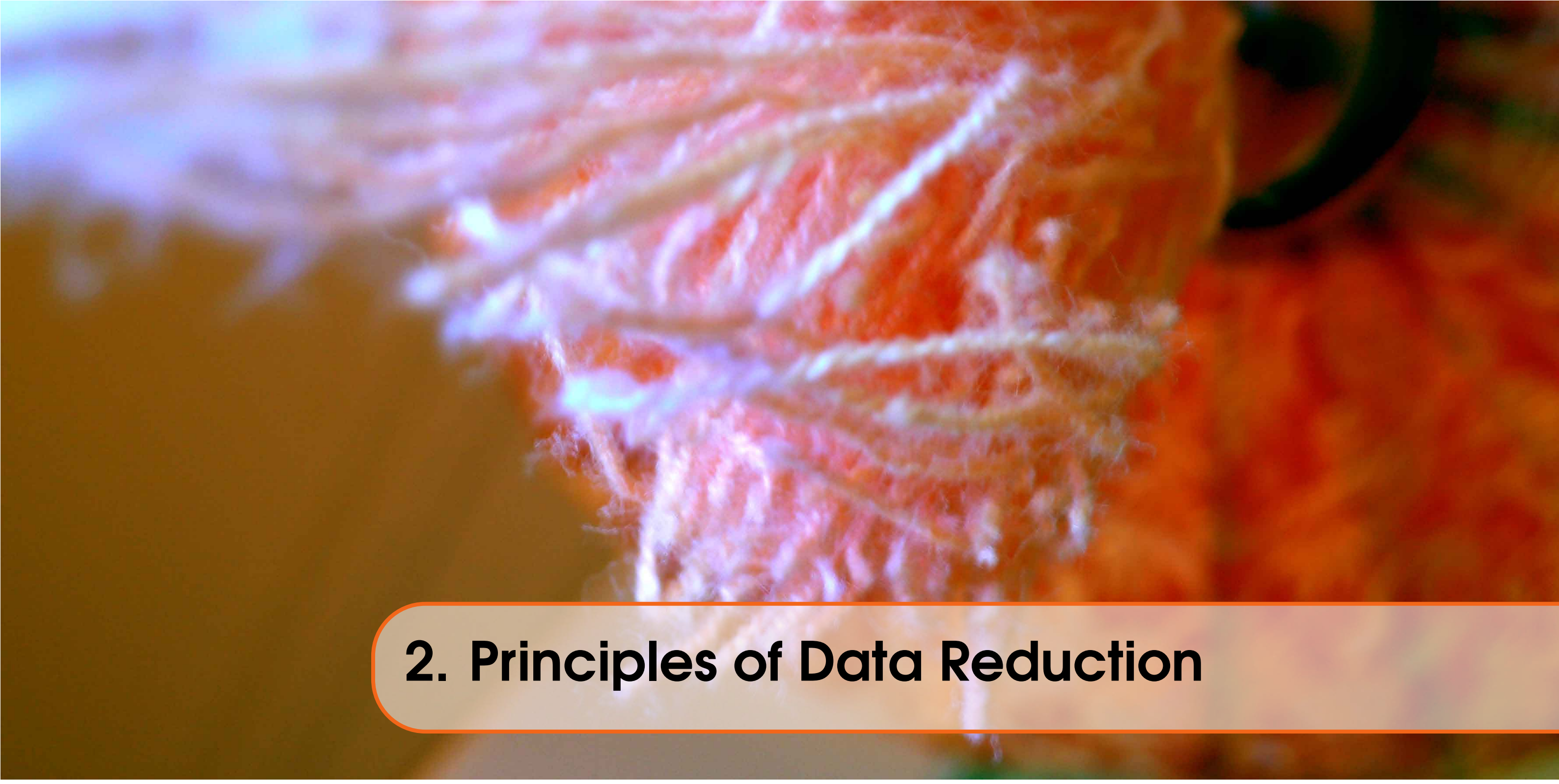
(b) Meh. You do it.
Problem 1.2 (C&B 7.6) Let $X_1, \ldots, X_n$ be a random sample from the pdf

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

(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, $\hat{\theta} = x_{(1)}$.



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.

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

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

for some g and h.

■ **Example 2.1** $X_1,\dots,X_n \overset{iid}{\sim}$ Bern(θ)
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

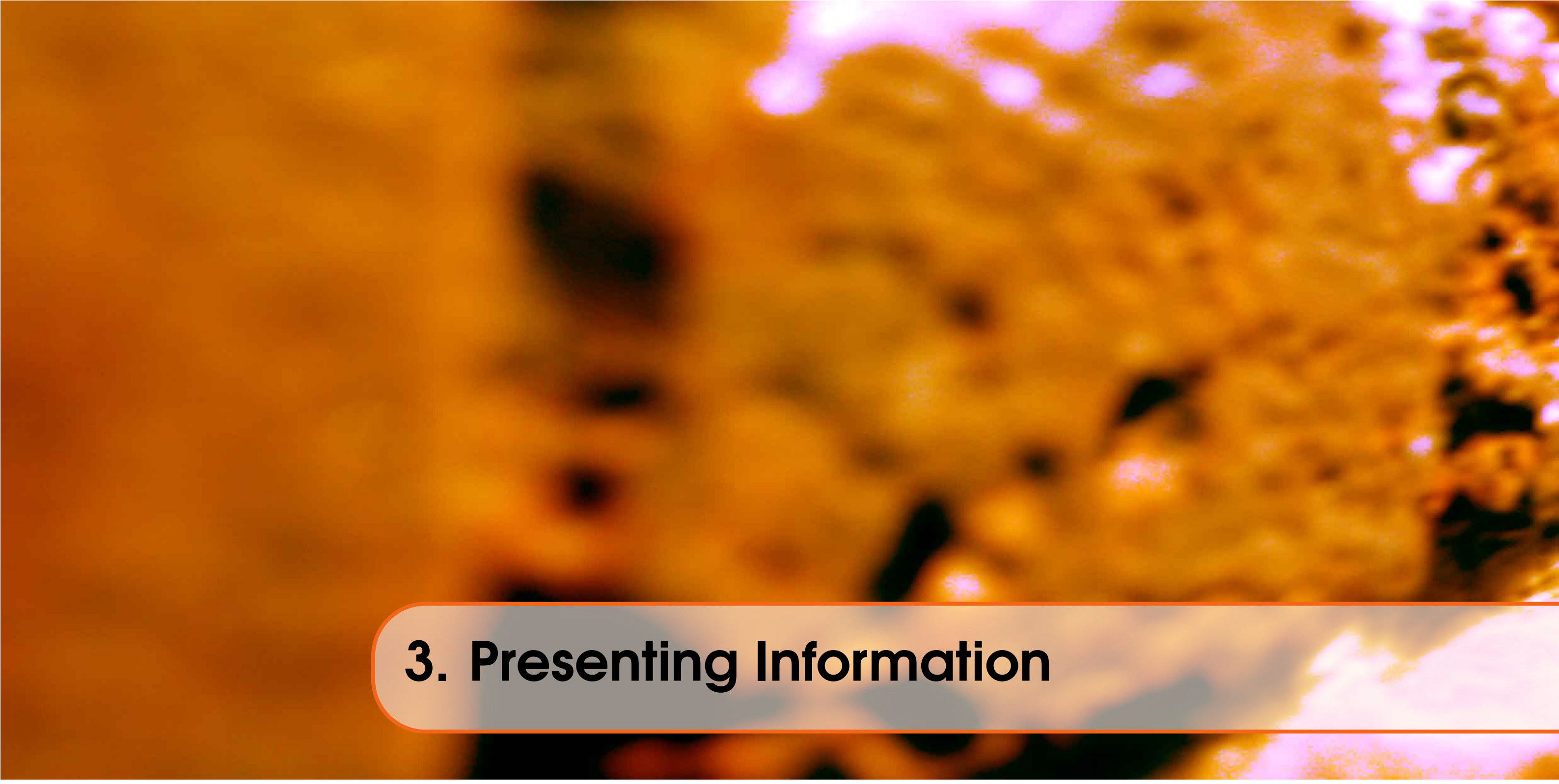
T(X) is sufficient

- is a function of any sufficient statistic

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$(\mu,\sigma^2)$.

$f(\underline{x}) = \prod(2\pi\sigma^2)^{\frac{1}{2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ ■



3. Presenting Information

3.1 Table

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table 3.1: Table caption

3.2 Figure

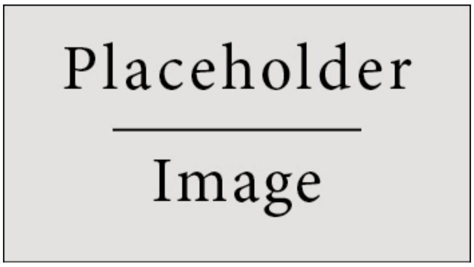


Figure 3.1: Figure caption

Bibliography

Books
Articles

Index

Figure, 15

Paragraphs of Text, 5

Table, 15