INTRODUCTION TO STATISTICAL INFERENCE

Sangita Das, notes by Ramdas Singh

Third Semester

List of Symbols

Placeholder

Contents

Chapter 1

SUFFICIENCY

1.1 Introduction to Sufficient Statistics

We start by defining terms for the sake of completion, whilst assuming the most basic definitions.

Definition 1.1. An *estimator* is any function of the random sample which is used to estimate the unknown value of the given paramteric function $g(\theta)$.

If $\underline{X} = (X_1, \ldots, X_n)$ is a random sample from a population with a probability distribution P_θ, a function $d(X)$ used for estimating $g(\theta)$ is known as an estimator. Let $\underline{x} = (x_1, \ldots, x_n)$ be a realization of $\underline{X} = (X_1, \ldots, X_n)$. Then $d(\underline{x})$ is called an *estimate*.

Definition 1.2. The *parameter space* is the set of all possible values of a parameter.

For example, the normal distribution $N(\mu, \sigma^2)$ has the parameter space $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Similarly, the binomial distribution $\text{Bin}(n, p)$ has the constraints $n \in \mathbb{N}$ and $p \in [0, 1]$.

Throughout this course, we will assume any data, otherwise stated, will be *independent and identically distributed*; the are separate datapoints that follow the same probability distribution and are indepedent.

Definition 1.3. Let $X_1, \ldots, X_n$ be a random sample from a population P_θ, where $\theta \in \Theta$. A statistic $T = T(X_1, \ldots, X_n) = T(\underline{X})$ is said to be a *sufficient statistic* for the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ is independent of θ.

We shall look at some examples.

Example 1.4. Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli distribution with parameter $p \in (0, 1)$. We claim that $T = \sum_{i=1}^n X_i$ is sufficient for $\{\text{Ber}(p) \mid 0 < p < 1\}$. To show this, we simply have

$$\begin{aligned} P(X_i = x_i \text{ for all } i | T = t) &= \frac{P(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n-1} x_i)}{P(\sum_{i=1}^n X_i = t)} \qquad (1.1) \\ &= \frac{P(X_1 = x_1) \cdots P(X_{n-1} = x_{n-1}) \cdot P(X_n = t - \sum_{i=1}^{n-1} x_i)}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{p^{x_1}(1-p)^{1-x_1} \cdots p^{x_{n-1}}(1-p)^{1-x_{n-1}} p^{t - \sum x_i}(1-p)^{1-t+\sum x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{1}{\binom{n}{t}}. \qquad (1.2) \end{aligned}$$

Thus, the statistic T is sufficient. The above expression is valid when $\sum_{i=1}^n x_i = t$, and the probability

evaluates to 0 if $\sum_{i=1}^n x_i \neq t$.

Example 1.5. Let $X_1, \ldots, X_n$ be a random sample from Poisson(λ) for $\lambda > 0$. We claim that the statistic $T = \sum_{i=1}^n X_i$ is sufficient. Recall that the probability mass function is $f(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$ where x is a non-negative integer, and $\lambda > 0$. We have

$$P(X_i = x_i \mid T = t) = \frac{P(X_1 = x_1) \cdots P(X_{n-1} = x_{n-1}) \cdot P(X_n = t - \sum_{i=1}^{n-1} x_i)}{P\left(\sum_{i=1}^n X_i = t\right)} \tag{1.3}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_{n-1}}}{x_{n-1}!} \cdot \frac{e^{-\lambda}\lambda^{t-\sum x_i}}{(t-\sum x_i)!}}{\frac{e^{-n\lambda}(n\lambda)^t}{t!}}$$

$$= \frac{e^{-n\lambda}\lambda^t}{x_1! \cdots x_{n-1}!(t - \sum_{i=1}^{n-1} x_i)!} \cdot \frac{t!}{e^{-n\lambda}(n\lambda)^t}$$

$$= \frac{t!}{x_1! \cdots x_{n-1}!\left(t - \sum_{i=1}^{n-1} x_i\right)!} \cdot \frac{1}{n^t}$$

$$= \binom{t}{x_1, x_2, \ldots, x_n} \cdot \frac{1}{n^t}. \tag{1.4}$$

This shows that the conditional distribution of $(X_1, \ldots, X_n)$ given $T = t$ does not depend on λ, so by the definition of sufficiency, T is a sufficient statistic for λ.

Definition 1.6. A *regular model* may be one of two things.

1. All P_θ are continuous with probability density function $f(x \mid \theta)$.
2. All P_θ are discrete with prbability mass function $p(x \mid \theta)$, and there exists a countable set $S = \{x_1, x_2, \ldots\}$ independent of θ such that $\sum_{i=1}^{\alpha} p(x_i|\theta) = 1$.

1.2 Factorization Theorems

The following theorem proves to be useful for finding sufficiency.

Theorem 1.7 (The *Neyman-Fisher factorization theorem*)**.** *Let $f(\underline{x} \mid \theta)$ be the density of $\underline{X}$ under the probability model P_θ for $\theta \in \Theta$. Then if the model is regular, a statistic $T(\underline{X})$ is sufficient for θ if and only if there exist functions g and h such that*

$$f(\underline{x} \mid \theta) = g(T(\underline{x}), \theta)h(\underline{x}). \tag{1.5}$$

Note that the functions are defined with $T : \mathbb{R}^n \to I \subseteq \mathbb{R}^k$ (for $k \leq n$), $g : I \times \Theta \to \mathbb{R}_{\geq 0}$, and $h : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$. The functions g and h need not be unique.

A little less formally, the theorem basically states this: let X be a random variable with probability mass/density function $f(x, \theta)$ for $\theta \in \Theta$. Then $T(X)$ is sufficient if and only if $f(x, \theta) = g(T(x), \theta)h(x)$ for all $\theta \in \Theta$. We now provide a proof.

Proof. We show only for the discrete case. Let us first assume such a faztorization exists. With

$$P_\theta(X = x' \mid T(X) = t) = \begin{cases} \frac{P_\theta(X=x', T(X)=t)}{P_\theta(T(X)=t)} & \text{if } T(x') = t, \\ 0 & \text{if } T(x') \neq t, \end{cases} \tag{1.6}$$

we then have

$$P_\theta(T(x) = t) = \sum_{\{x|T(x)=t\}} f_\theta(x \mid \theta) = g(T(x), \theta) \sum_{\{x|T(x)=t\}} h(x). \tag{1.7}$$

Thus, using the above, and the fact that $\{X = x\} \subseteq \{T(X) = T(x)\}$, gives us

$$\frac{P_\theta(X = x', T(X) = t)}{P_\theta(T(X) = t)} = \frac{P_\theta(X = x')}{g(T(x), \theta) \sum_{\{x|T(x)=t\}} h(x)} = \frac{g(t, \theta)h(x')}{g(T(x), \theta) \sum_{\{x|T(x)=t\}} h(x)} = \frac{h(x')}{\sum_{\{x|T(x)=t\}} h(x)}. \tag{1.8}$$

We now suppose that $T(X)$ is sufficient for θ. Let $g(t, \theta) = P_\theta(T = t)$. Then,

$$g(t, \theta) = P_\theta(T = t) = P_\theta(T(X) = T(x')) \text{ where } T(x') = t. \tag{1.9}$$

Also set $h(x) = P_\theta(X = x' \mid T(X) = T(x'))$, which is independent of θ since T is sufficient. Therefore, we have

$$f_X(x' \mid \theta) = P_\theta(X = x') = P_\theta(T(X) = T(x')) \cdot P_\theta(X = x' \mid T(X) = T(x')) = g(T(x), \theta)h(x). \tag{1.10}$$

■

Example 1.8. Let $X_1, \ldots, X_n$ be independent and identically distributed $N(\mu, \sigma^2)$ random variables, with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Let us find a sufficient test statistic. We look at cases; the first case being when σ^2 is known ($\sigma^2 = 1$). Since these are independent, we have the joint probability density funciton of these random variables as

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^{n}(x_i - \mu)^2} \tag{1.11}$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right)\right)$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \times e^{-\frac{1}{2}(-2\mu \sum_{i=1}^{n} x_i + n\mu^2)}. \tag{1.12}$$

Make the former term $h(x)$ and the latter term $g(\sum_{i=1}^n x_i, \mu)$ with $T(x) = \sum_{i=1}^n x_i$. The second case now involves μ being known, and we set it to $\mu = 0$ to get $T(x) = \sum_{i=1}^n x_i^2$, $h(x) = 1/(2\pi)^{n/2}$, and $g(T(x), \sigma^2) = \sigma^{-n} e^{-T(x)/2\sigma^2}$.

We move on to another factorization theorem.

Definition 1.9. The family of distributions $\{P_\theta \mid \theta \in \Theta\}$ is said ot be a *single parameter exponential family* if there eixst real valued functions $c(\theta), d(\theta)$ on Θ and $T(x), S(x)$ on $\mathbb{R}^n$ and a set $A \subset \mathbb{R}^n$ such that

$$f(\underline{x} \mid \theta) = \exp(c(\theta)T(\underline{x}) + d(\theta) + S(x))\mathbf{1}_A(x) \tag{1.13}$$

where A must not depend on θ.

Example 1.10. Suppose $X \sim \text{Poisson}(\lambda)$ for $\lambda > 0$. With $A = \{0, 1, 2, \ldots\}$, we have

$$f(x \mid \lambda) = \exp(x \log(\lambda) - \lambda - \log(x!))\mathbf{1}_A(x_)) \tag{1.14}$$

with $T(x) = x$, $c(\lambda) = \log(\lambda)$, $d(\lambda) = -\lambda$, and $S(x) = -\log(x!)$.

Consider $X_1, \ldots, X_n$ independent and identically distributed random variables following the distribution P_θ, and suppose that $\{P_\theta \mid \theta \in \Theta\}$ is an exponential family, that is, $f(x \mid \theta) = \exp(c(\theta)T(x_i) +$

$d(\theta) + S(x))\mathbf{1}_A(x)$. Then,

$$f_{x_1,\ldots,x_n}(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} \exp(c(\theta)T(x_i) + d(\theta) + S(x_i))\mathbf{1}_A(x_i) \tag{1.15}$$

$$= \exp(c(\theta)\sum_{i=1}^{n} T(x_i) + md(\theta) + \sum_{i=1}^{n} S(x_i))\mathbf{1}_{A^n}(x_1,\ldots,x_n). \tag{1.16}$$

$(x_1,\ldots,x_n)$ has distribution belonging to a single parameter exponential family. Thus, if $\{P_\theta \mid \theta \in \Theta\}$ is a single parameter family with density $f(x,\theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))\mathbf{1}_A(x)$, then $T(x)$ is sufficient for θ.

Corollary 1.11. *If $x_1,\ldots,x_n$ are independent and indentically distributed random variables following the distribution P_θ with density $f(x \mid \theta) = \exp(c(\theta)T(x) + d(\theta) + S(x))\mathbf{1}_A(x)$, then $\sum_{i=1}^{n} T(X_i)$ is sufficient for θ.*

The exponential family is expanded.

Definition 1.12. A family of distributions $\{P_\theta : \theta \in \Theta\}$ with density $f(x \mid \theta)$ is called a *k-parameter exponential family* if there exists real valued functions $c_1(\theta),\ldots,c_k(\theta), d(\theta)$ on Θ and $T_1(\underline{x}),\ldots,T_k(\underline{x}), S(x)$ on $\mathbb{R}^n$, and a set $A \subset \mathbb{R}^n$ such that

$$f(\underline{x} \mid \theta) = \left(\exp(\sum_{j=1}^{n} c_j(\theta)T_j(\underline{x}) + d(\theta) + S(\underline{x}))\right)\mathbf{1}_A(\underline{x}). \tag{1.17}$$

Here, $(T_1,\ldots,T_k)$ is a k-dimensional sufficient statistic for θ. Note that the parameter here is θ and not $(c_1(\theta),\ldots,c_k(\theta))$.

We look at more examples.

Example 1.13. For a normal distribution with $\sigma^2 = 1$, we have

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \mathbf{1}_A(x) = \exp\left(-\frac{1}{2}\log(2\pi) - \frac{x^2}{2} + x\theta - \frac{\theta^2}{2}\right)\mathbf{1}_A(x). \tag{1.18}$$

Here, $c(\theta) = \theta$, $T(x) = x$, $S(x) = -\frac{x^2}{2} - \frac{1}{2}\log(2\pi)$, and $d(\theta) = -\frac{\theta^2}{2}$.

August 1st.

Remark 1.14.

1. The Neyman-Fisher factorization theorem holds if $\underline{\theta}$ and $\underline{T}$ are vectors. Their dimensions need not be equal.
2. If T is sufficient and T is a function of U, then U is also sufficient.
3. If V is a function of T, then V need not be sufficient. But if V is one-to-one with T, then V is also sufficient. $V = B(T)$ and $T = B^{-1}(V)$ shows that $g(T,\theta) = g(B^{-1}(V),\theta) = g^*(V,\theta)$. Note that the inverse exists since it is defined on the image of the original function only.

1.3 Minimal Sufficiency

Again, we being with a few definitions.

Definition 1.15. A *partition* of a space $\mathcal{X}$ is a collection $\{E_i\}$ of subsets of $\mathcal{X}$ such that

$$\bigcup_{n\geq 1} E_i = \mathcal{X} \text{ and } E_i \cap E_j = \emptyset \text{ for } i \neq j. \tag{1.19}$$

The E_i's are called *partition sets*. Let $T : \mathcal{X} \to \mathcal{Y}$. The partition of $\mathcal{X}$ induced by the function T is the collection of the sets $T_y = \{x \mid T(x) = y\}$ for $y \in \mathcal{Y}$.

We say that $\mathcal{P}_2$ is a *reduction* of $\mathcal{P}_1$ if each partition set of $\mathcal{P}_2$ is the union of the same members of $\mathcal{P}_1$.

Definition 1.16. A sufficient statistic $T(X)$ is called a *minimal sufficient statistic* if for any other sufficient statistic $T'(X)$, $T(\underline{X})$ is a function of $T'(X)$. That is,

$$T(\underline{X}) = U(T'(X)) \implies \text{ if } T'(\underline{x}) = T'(\underline{y}) \text{ then } T(\underline{x}) = T(\underline{y}). \tag{1.20}$$

In terms of partition sets, if $\{B_{t'} \mid t' \in T'\}$ are partition sets for $T'(x)$ and $\{A_t : t \in T\}$ are partition sets for $T(x)$, then the definition states that every $B_{t'}$ is a subset of some A_t. Thus the partition associated with a minimal sufficient statistic is the coarsest possible partition for a sufficient statistic, and a minimal sufficient statistic achieves the greatest possible data reduction.

Theorem 1.17. *Let $f(x \mid \theta)$ be the probability mass/density funciton of a sample $\underline{X}$. Suppose there exists a function $T(\underline{x})$ such that for every two sample points $\underline{x}$ and $\underline{y}$, the ratio $f(\underline{x} \mid \theta)/f(\underline{y} \mid \theta)$ is constant as a function of θ if and only if $T(\underline{x}) = T(\underline{y})$. Then $T(\underline{X})$ is a minimal sufficient statistic for θ.*

We look at an example first before proving the theorem.

Example 1.18. Let $X_1, \ldots, X_n$ be independent and identically distributed Exp(θ) for $\theta > 0$. Recall that the probabilty density function is $f(x \mid \theta) = \theta \exp(-\theta x)$. We show that $T(\underline{X}) = \sum_{i=1}^n X_i$ is minimal sufficient for θ. The joint density in this case is

$$f(\underline{X} = \underline{x} \mid \theta) = \prod_{i=1}^n \theta \exp(-\theta x_i) = \theta^n \exp\left(-\theta . \sum_{i=1}^n x_i\right). \tag{1.21}$$

The ratio is now

$$\frac{f(\underline{x} \mid \theta)}{f(\underline{y} \mid \theta)} = \exp\left(-\theta \sum_{i=1}^n (x_i - y_i)\right) = \exp(-\theta(T(\underline{x}) - T(\underline{y}))). \tag{1.22}$$

This expression is constant as a function of θ if and only if $T(\underline{x}) = T(\underline{y})$. Thus, T is minimal sufficient statistic for θ.

Proof. We shall assume that $f(x \mid \theta) > 0$ for all $x \in X, \theta \in \Theta$. Suppose there exists $T(X)$ such that $f(\underline{x} \mid \theta)/f(\underline{y} \mid \theta)$ is constant as a function of θ if and only if $T(x) = T(y)$. We first show that T is sufficient. The map is really $T : \mathcal{X} \to \mathcal{T} = \{t \mid T(x) = t \text{ for some } x \in \mathcal{X}\}$. Let $A_t = \{x \in \mathcal{X} \mid T(x) = t\}$. Then the collection of sets $\{A_t\}_{t\in\mathcal{T}}$ is a partition of $\mathcal{X}$.

For each A_t, fix an element $x_t \in A_t$. For any $x \in \mathcal{X}$, we have $x \in A_{T(x)}$ and hence $x_{T(x)}$ is the fixed element which belongs to the same partitioning set as x does. Thus, $T(x) = T(x_{T(x)})$ since x and $x_{T(x)}$ belong to $A_{T(x)}$. $\frac{f(x|\theta)}{f(x_{T(x)}|\theta)}$ is a constant function of θ, so $h(x) = \frac{f(x|\theta)}{f(x_{T(x)}|\theta)}$ independent of θ and $h : \mathcal{X} \to \mathbb{R}_{\geq 0}$. Define $g : \mathcal{T} \times \Theta \to \mathbb{R}_{\geq 0}$ by $g(t, \theta) = f(x_t \mid \theta)$. Then

$$f(x \mid \theta) = \frac{f(x \mid \theta)}{f(x_{T(x)} \mid \theta)} f(x_t \mid \theta) = h(x)g(t, \theta). \tag{1.23}$$

Now that we have shown T is sufficient, we show its minimality. Let $T'(X)$ be any other sufficient statistic. Then there exist functions g' and h' such that

$$f(x \mid \theta) = g'(T'(x), \theta)h'(x). \tag{1.24}$$

Let x and y be any two sample points such that $T'(x) = T'(y)$. Then

$$\frac{f(x \mid \theta)}{f(y \mid \theta)} = \frac{g'(T'(x), \theta)h'(x)}{g'(T'(y), \theta)h'(y)} = \frac{h'(x)}{h'(y)} \text{ is independent of } \theta. \tag{1.25}$$

We already know that $T(x) = T(y)$ whenever the above ratio is a constant function of θ. Hence, $T'(x) = T'(y) \implies T(x) = T(y)$. This means that T is coarser. ■

Theorem 1.19. *Suppose $\mathcal{P}$ is a family of probability models with common support and $\mathcal{P}_0 \subset \mathcal{P}$. If T is minimal sufficient for $\mathcal{P}_0$ and sufficient for $\mathcal{P}$, then it is minimal sufficient for $\mathcal{P}$ also.*

Proof. Let U be any sufficient statistic for $\mathcal{P}$. Then it is sufficient for $\mathcal{P}_0$. But T is minimal for $\mathcal{P}_0$. Therefore, $T = H(U)$. Now consider $\mathcal{P}$. T is sufficient for $\mathcal{P}$ and for any other sufficient statistic U, $T = H(U)$. Thus, T is minimal sufficient. ■

Example 1.20. Let $X_1, \ldots, X_n$ be independent and identically distributed Poisson(λ) random variables. The probability mass function in this case is

$$f(x_1, \ldots, x_n \mid \lambda) = e^{-n\lambda} \frac{\lambda^{x_1 + \cdots + x_n}}{x_1! \cdots x_n!}. \tag{1.26}$$

We find whether $\sum_{i=1}^n X_i$ is sufficient for λ. We have

$$\frac{f(\underline{x} \mid \theta)}{f(\underline{y} \mid \theta)} = \theta^{-(\sum_{i=1}^n x_i - \sum_{j=1}^n y_j)} \frac{y_1! \cdots y_n!}{x_1! \cdots x_n!} \tag{1.27}$$

which is a constant with respect to θ if and only if $T(\underline{x}) = T(\underline{y})$.

Definition 1.21. Two statistics S_1 and S_2 are said to be *equivalent statistics* if $S_1(x) = S_1(y)$ if and only if $S_2(x) = S_2(y)$. Note that if S_1 and S_2 are equivalent, then then provide the same

1. partition of the sample space,
2. reduction, and
3. information.

Definition 1.22. A statistic $S(\underline{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic*. An example is the chi-squared distribution.

1.4 Location Scale Family

With examples as context, we define the following families.

Example 1.23. Consider $U \sim \text{Unif}(-1, 1)$. Then $f_U(u) = \frac{1}{2} I_{(-1,1)}(u)$. Let $X = \mu + U$. Then $X \sim \text{Unif}(\mu - 1, \mu + 1)$. Thus,

$$f_X(x) = \frac{1}{2} I_{(\mu-1,\mu+1)}(x) = \frac{1}{2} I_{(-1,1)}(x - \mu) = f_U(x - \mu). \tag{1.28}$$

The family of distributions for X indexed by μ is called a *location family* with *location parameter* μ. Note that μ is the location for X if $X - \mu$ has a distribution which is free of μ.

Example 1.24. Suppose $Z_1 \sim N(0,1)$ with density $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$. If we set $X = \sigma Z$ with $\sigma > 0$, then $X \sim (0, \sigma^2)$. Thus,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{\sigma} f_Z(\frac{x}{\sigma}). \tag{1.29}$$

Here, σ is called the *scale parameter* for the family of distributions X indexed by it, which is called a *scale family*. Together, we have the changed distribution as

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right). \tag{1.30}$$

Chapter 2

POINT ESTIMATION

August 8th.

We begin with a definition.

Definition 2.1. A *point estimator* is a function W of the sample, mapping into the parameter space Θ of the parameter θ of interest. The value $W(X)$ is called a *point estimate* of θ.

Various methods of point estimation are discussed in this chapter, including the method of moments, maximum likelihood estimation, and Bayes estimation.

2.1 Estimators

2.1.1 Method of Moments

Let $X_1, \ldots, X_n$ be a sample from a population with a probability distribution function and probability density function. The *method of moments* estimators are found by equating the first k sample moments to the corresponding k population moments and solving the resulting system of simultaneous equations. Here, the k^{th} sample moment and the k^{th} population moment are given as

$$m_k = \frac{1}{n}\sum_{j=1}^{n} X_j^k \qquad \text{and} \qquad \mu_k' = E[X^k] \tag{2.1}$$

respectively. For $\mu_j' = \mu_j'(\theta_1, \ldots, \theta_i)$, with $1 \leq j \leq k$, the estimators are obtained by solving the equations

$$m_j = \mu_j'(\theta_1, \ldots, \theta_i) \qquad \text{for } 1 \leq j \leq k. \tag{2.2}$$

2.1.2 Maximum Likelihood Estimators

Let $X_1, \ldots, X_n$ be an independent and identically distributed sample from a population with a probability distribution/mass function $f(\underline{x} \mid \theta_1, \ldots, \theta_k)$. The *likelihood function* is defin as

$$L(\underline{\theta} \mid \underline{x}) = L(\theta_1, \ldots, \theta_k \mid x_1, \ldots, x_n) := \prod_{j=1}^{n} f(x_j \mid \theta_1, \ldots, \theta_k). \tag{2.3}$$

Definition 2.2. For each sample points, let $\overline{\theta}(\underline{x})$ be a parameter value at which $L(\underline{\theta} \mid \underline{x})$ attains its maxima as a function of $\underline{\theta}$, with $\underline{x}$ held fixed. A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample $\underline{X}$ is $\hat{\theta}(\underline{X})$.

Remark 2.3. 1. By its contraction, the ranges of the MLE concludes with the range of the parameter.

2. The MLE is the parameter points for which the observation sample is most likely.

If the likelihood function is differentiable in θ_i, then possible candidates for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that solve $\frac{\partial}{\partial \theta_i} L(\underline{\theta} \mid \underline{x}) = 0$ for $1 \leq i \leq k$. Moreover, we must have $\frac{\partial^2}{\partial^2 \theta_i} L(\underline{\theta} \mid \underline{x}) < 0$ for the solution to be a maximum.

Remark 2.4. 1. The solutions to the above equation are the only possible candidates for the MLE since the first derivative being zero is only a necessary condition for the maximum and not sufficient.

2. The zeroes of the first derivatives only locate extreme points in the interior of the domain of the function.

3. If the extrema occurs on te boundary of the domain, then the first derivative may not be zero. Thus the boundary must be checked separately.

4. The points where the first derivatives are zero may be local/global maxima or minima.

Example 2.5. Let us take the example of the binomial distribution, $X \sim \text{Bin}(n, p)$. The likelihood function is given by $L(p \mid x) = \binom{n}{x} p^x (1-p)^{n-x}$, where $0 < p < 1$. We have

$$\frac{\partial}{\partial p} L(p \mid x) = \frac{dL}{dp}(p \mid x) = 0. \tag{2.4}$$

This derivative is hard to compute directly. So we take the logarithm (an increasing function) of the likelihood function, and then differentiate.

$$\implies \frac{d}{dp} \log L(p \mid x) = \frac{x}{p} - \frac{n-x}{1-p} = 0. \tag{2.5}$$

This gives us $p = \frac{x}{n}$, with the second derivative less than zero, confirming that this is a maximum. Thus, the MLE of p is $\hat{p} = \frac{X}{n}$.

The method in this example is known as *log likelihood estimation*. It is often easier to work with the log likelihood function, especially when dealing with products.

Remark 2.6. 1. The MLE may not exist at all or may not be unique.

2. If $\hat{\theta}$ is the MLE of θ, then $\rho(\hat{\theta})$ is the MLE of $\rho(\theta)$ for any function ρ.

The following is a vital and important result.

Theorem 2.7. *The MLE depends on $\underline{x}$ only through the sufficient statistic $T(\underline{x})$.*

Proof. We have $L(\theta \mid \underline{x}) = f(\underline{x} \mid \theta) = g(T(\underline{x}), \theta) h(\underline{x})$. Therefore, we have

$$L(\hat{\theta}(\underline{x}) \mid \underline{x}) = \max_{\theta} g(T(\underline{x}), \theta) h(\underline{x}). \tag{2.6}$$

Since $h(\underline{x}) > 0$, and does not depend on θ, we must have

$$L(\hat{\theta}(\underline{x}) \mid \underline{x}) = h(\underline{x}) \max_{\theta} g(T(\underline{x}), \theta) \tag{2.7}$$

where the maximization is on the part that involves $\underline{x}$ through $T(\underline{x})$ only. ∎

2.1.3 One Parameter Exponential Family in Natural Form

Recall that the usual density form of the exponential family was given as

$$f(x \mid \theta) = \exp(c(\theta)T(x) + d(\theta) + s(x))\mathbf{1}_A(x). \tag{2.8}$$

Define $\eta = c(\theta)$ for $\theta \in \Theta$, and let $\Gamma = \{\eta \mid \eta = c(\theta), \theta \in \Theta\}$. We then have

$$f^*(x \mid \eta) = \exp(\eta T(x) + d_0(\eta) + s(x))\mathbf{1}_A(x) \tag{2.9}$$

where $d_0(\eta) = d(c^{-1}(\eta))$ if c is one-one. Moreover,

$$1 = \int_A f^*(x \mid \eta)dx = \int_A \exp(\eta T(x) + d_0(\eta) + s(x))dx = \exp(d_0(\eta)) \int_A \exp(\eta T(x) + s(x))dx \tag{2.10}$$

$$\implies d_0(\eta) = -\log\left(\int_A \exp(\eta T(x) + s(x))dx\right). \tag{2.11}$$

Theorem 2.8. *If X has density of the form $f(x \mid \eta) = \exp(\eta T(x) + d_0(\eta) + s(x))\mathbf{1}_A(x)$ and η is an interior points of $H := \{\eta \mid |d_0(\eta)| < \infty\}$, then the moment generating function of $T(X)$ exists and is given by*

$$\varphi(s) = E[\exp(sT(X))]. \tag{2.12}$$

Also,

$$E[T(X)] = -\frac{d}{d\eta}d_0(\eta), \qquad \text{and} \qquad \text{Var}(T(X)) = -\frac{d^2}{d\eta^2}d_0(\eta). \tag{2.13}$$

Theorem 2.9. *Let $\{P_\theta \mid \theta \in \Theta\}$ be a one parameter exponential family with density $f(x \mid \theta) = \exp(c(\theta)T(x) + d(\theta) + s(x))\mathbf{1}_A(x)$ and let c be the interior of $C = \{c(\theta) \mid \theta \in \Theta\}$. Also suppose $\theta \mapsto c(\theta)$ is one-one. If the equation*

$$E_\theta[T(X)] = T(x) \tag{2.14}$$

has a solution $\hat{\theta}(x)$ for which $c(\hat{\theta}(x)) \in C$, then $\hat{\theta}(x)$ is the unique MLE of θ.

Proof. Since $\theta \mapsto c(\theta)$ is one-one, maximizing the likelihood over θ is maximizing over $\eta = c(\theta)$. Hence, consider the natural parametrization

$$f(x \mid \eta) = \exp(\eta T(x) + d_0(\eta) + s(x))\mathbf{1}_A(x) \text{ for } \eta \in H. \tag{2.15}$$

$L(\eta \mid x) = \eta T(x) + d_0(\eta) + s(x)$, if $x \in A$, is the log likelihood function. Also,

$$\frac{\partial}{\partial\eta}L(\eta \mid x) = T(x) + d_0'(\eta) \text{ and } \frac{\partial^2}{\partial\eta^2}L(n \mid x) = d_0''(\eta). \tag{2.16}$$

Therefore, we get $\frac{\partial}{\partial\eta}L(\eta \mid x) = T(x) - E_\eta[T(x)] = 0$ implying that $E_\eta[T(X)] = T(x)$. Now, $\frac{\partial^2}{\partial\eta^2}L(\eta \mid x) < 0$ so that L is strictly concave. Then we get a unique maxima at $\hat{\eta}(x)$ for which $E_\eta[T(X)]_{\eta=\hat{\eta}(x)} = T(x)$. ■

August 22nd.

The *loss function* $L(q(\theta), T(X))$ measures the discrepancy between the true parameter $q(\theta)$ and the estimate $T(X)$. A common choice for L is the squared error loss, defined as

$$L(q(\theta), T(X)) = (q(\theta) - T(X))^2. \tag{2.17}$$

One may also choose the absolute error loss, defined as

$$L(q(\theta), T(X)) = |q(\theta) - T(X)|. \tag{2.18}$$

Theorem 2.10 (The *Rao-Blackwell theorem*)**.** *Let $\underline{X}$ be a random vector with distribution P_θ, $\theta \in \Theta$, and let T be sufficient for θ. Moreover, let $\delta(\underline{X})$ be an estimator of θ and $\delta^*(t) = E[\delta(X) \mid T = t]$. Also let $L(\theta, d)$ be a strictly convex loss function (in d) and $R(\theta, d) = E[L(\theta, d(\underline{X}))]$. Then if $R(\theta, \delta) = E[L(\theta, \delta(X))] < \infty$, we obtain $R(\theta, \delta^*) < R(\theta, \delta)$ for all θ unless $\delta(x) = \delta^*(T(x))$ with probability 1.*

Another form is

Theorem 2.11. *If T is an unbiased estimate of $\tau(\theta)$ and S is a sufficient statistic, then $T' = E_\theta[T \mid S]$ is also unbiased for $\tau(\theta)$ and*

$$\mathrm{Var}_\theta(T') \leq \mathrm{Var}_\theta(T) \text{ for all } \theta. \tag{2.19}$$

Proof. We simply have $E_\theta[T'] = E_\theta[E[T \mid S]] = \tau(\theta) = E_\theta[T] = \tau(\theta)$, where we have used the property $E[E[X \mid Y]] = E[X]$. For the next part, we have

$$\begin{aligned}\mathrm{Var}_\theta(T) &= E[T - E[T]]^2 = E[T - \tau(\theta)]^2 = E[T - T' + T' - \tau(\theta)]^2 \\ &= E[T - T']^2 + E[T' - \tau(\theta)]^2 + 2E[(T - T')(T' - \tau(\theta))].\end{aligned} \tag{2.20}$$

The last term can be worked upon as

$$E[(T - T')(T' - \tau(\theta))] = E[E[(T - T')(T' - \tau(\theta)) \mid S]] = E[(T - \tau(\theta))E[(T - T') \mid S]] \tag{2.21}$$

where we have used the fact that $E[AB \mid S] = AE[B \mid S]$ when B is any random variable and A is measurable with respect ot S. But note that the inner term, $E[(T - T') \mid S]$ expands as

$$E[(T - T') \mid S] = E[T \mid S] - E[T' \mid S] = 0 \tag{2.22}$$

giving us

$$\mathrm{Var}_\theta(T) = E[T - T']^2 + E[T' - \tau(\theta)]^2 \geq E[T' - \tau(\theta)]^2 = \mathrm{Var}_\theta(T'). \tag{2.23}$$

■

Theorem 2.12. *Let X have distribution P_θ, $\theta \in \Theta$, and let $T = T(\underline{X})$ be complete and sufficient for θ (or P_θ, $\theta \in \Theta$). Then every function $h(T)$ is the unique unbiased estimator of its own expected value; that is, for any h, if $g(\theta) = E_\theta[h(T)]$ holds, then $h(T)$ is the only unbiased estimate available for $g(\theta)$.*

Proof. T is complete for θ, so for any function h, $E[h(T(X))] = 0$ for all θ implies that $h \equiv 0$. Let $h_1(T)$ and $h_2(T)$ be unbiased, so that $E[h_1(T)] = \theta = E[h_2(T)]$. Set $h(t) = h_1(t) - h_2(t)$. Then $E[h(T)] = 0$ for all θ, which implies that $h \equiv 0$, or $h_1(t) = h_2(t)$. ■

Theorem 2.13 (The *Lehmann-Scheffe theorem*)**.** *Suppose $T = T(\underline{X})$ is complete sufficient for P_θ, $\theta \in \Theta$, and $S = S(\underline{X})$ is any unbiased estimate of $q(\theta)$. Then $S^* = E[S(\underline{X}) \mid T]$ is the unique minimum variance unbiased estimator (UMVUE) of $q(\theta)$ is $\mathrm{Var}_\theta(S^*(\underline{X})) < \infty$ for all θ.*

Proof. Both S and S^* are unbiased, and the mean squared error is the variance. By Rao-Blackwell theorem, $\mathrm{Var}_\theta(S^*) \leq \mathrm{Var}_\theta(S)$ for all θ. Let S_1 and S_2 be two such that $g_1(T) = E[S_1 \mid T]$ and $g_2(T) = E[S_2 \mid T]$ with $E[g_1(T)] = \theta = E[g_2(T)]$. Then $E[h(T)] = E[g_1(T) - g_2(T)] = 0$ showing $h \equiv 0$. ■

Remark 2.14.

1. Give any $S(X)$ unbiased for $q(\theta)$, the UMVUE is found by obtaining $S^*(X) = E[S(X) \mid T(X)]$, where T is a complete sufficient statistic for θ.
2. If we already have $h(T)$ unbiased for $q(\theta)$ and T complete sufficient, then $h(T)$ is the UMVUE for $q(\theta)$ since $S^* = E[h(T) \mid T] = h(T)$.

We work with some problems to familiarize.

Example 2.15. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be independent and identically distributed random variables. Setting $s^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$, one can show that s^2 is unbiased for σ^2.

Example 2.16. For $X_1, \ldots, X_n \sim N(\theta, \theta^2)$, one can show that $(\sum X_i, \sum X_i^2)$ is sufficient but not complete.

Theorem 2.17 (*Basu's theorem*)**.** *Suppose T is complete sufficient for $\{P_\theta \mid \theta \in \Theta\}$. Let S be any ancillary statistic. Then T and S are indpendent for all θ.*

Proof. We have

$$\int_{-\infty}^{\infty} [F_S(s) - F_{S|T=t}(s)] f_T(t) dt = 0 \text{ for all } \theta. \tag{2.24}$$

■

Theorem 2.18. *Let $\{P_\theta \mid \theta \in \Theta\}$ be a k-parameter exponential family with density*

$$f(\underline{x} \mid \theta) = \exp\left(\sum_{j=1}^{k} c_j(\theta) T_j(\underline{x}) + d(\theta) + S(\underline{x})\right) \mathbf{1}_A(\underline{x}). \tag{2.25}$$

Suppose $\{\underline{c}(\theta) = (c_1(\theta), \ldots, c_n(\theta)) \mid \theta \in \Theta\}$ contains an open set $g(e)$ in $\mathbb{R}^k$. Then $\underline{T}(\underline{X}) = (T_1, \ldots, T_k)$ is complete sufficient.

Theorem 2.19. *A (bounded) complete sufficient statistic is minimal sufficient, assuming that the minimal sufficient statistic exists.*

Proof. Let T be minimal sufficient and U complete sufficient.Then $T = h(U)$ for some function h. We need to show that T and U are equivalent statistics (produce the same partition). It is enough t show that for all integrable φ, $E[\varphi(U) \mid T] = \varphi(U)$. Suppose $E[\varphi(U) \mid T] \neq \varphi(U)$ for some φ. Define $K(U) = \varphi(U) - E[\varphi(U) \mid h(U)]$. Then

$$E[K(U)] = E[\varphi(U)] - E[E[\varphi(U) \mid h(U)]] = E[\varphi(U)] - E[\varphi(U)] = 0. \tag{2.26}$$

But U is complete, so $K(U) \equiv 0$, or $\varphi(U) = E[\varphi(U) \mid h(U)]$ for all integrable φ. This shows that T and U are equivalent statistics. ■

2.2 Information Number

Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability distributions satisfying the following mathematical regularity conditions.

(A) $A = \{x : f(x) > 0\}$ does not depend on θ.

For all $x \in A$, $\theta \in \Theta$, the *score function*

$$S(x) = \frac{\partial}{\partial \theta} \log f(x \mid \theta) = \frac{\frac{\partial}{\partial \theta} f(x \mid \theta)}{f(x \mid \theta)} \tag{2.27}$$

exists and is finite. $S(x)$ measures the relative rate at which $f(x \mid \theta)$ changes at x. Since X is random, this needs averaging as

$$I(\theta) = E_\theta[S(X)^2] = \int \left(\frac{\partial}{\partial \theta} \log f(x \mid \theta)\right)^2 f(x \mid \theta) dx. \tag{2.28}$$

$I(\theta)$ is called the *Fisher information number* about θ contained in the observation X. It measures the amount of information that the observable random variable X carries about the unknown parameter θ. The second regularity condition is

(B) The derivative, with respect to θ, of $\int f(x \mid \theta)dx$ can be obtained by differentiating under the integral sign.

Theorem 2.20. *If both (A) and (B) hold then*

1. $E_\theta[S(X)] = 0$,
2. $I(\theta) = \mathrm{Var}_\theta(S(X))$.

In addition, if the second derivative (w.r.t. θ) of $\log f(x \mid \theta)$ for all x and θ, and the second derivative of $\int f(x \mid \theta)dx$ can be obtained by differentiating under the integral sign, then

3. $I(\theta) = -E_\theta[\frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta)]$.

Theorem 2.21. *Let X, Y be two independent random variables with densities $f_1(x \mid \theta)$ and $f_2(y \mid \theta)$ respectively. If $I(\theta)$, $I_1(\theta)$, $I_2(\theta)$ are the Fisher information numbers of (X, Y), X, Y respectively, then $I(\theta) = I_1(\theta) + I_2(\theta)$.*

The above theory works for single parameters; for multiple parameters, the *fisher information matrix* is defined as

$$I(\theta) = [I_{ij}(\theta)] \text{ where } I_{ij}(\theta) = E\left[\frac{\partial}{\partial\theta_i} \log f(x \mid \theta) \cdot \frac{\partial}{\partial\theta_j} \log f(x \mid \theta)\right]. \tag{2.29}$$

The matrix is symmetric.

August 29th.

Theorem 2.22. *Let X have one parameter exponential family density*

$$f(x \mid \theta) = \exp(c(\theta)T(x) + d(\theta) + s(x))\mathbf{1}_A(x). \tag{2.30}$$

Consider the mean value parametrization $\delta(\theta) = E_\theta[T(X)]$. Then $I(\delta) = \frac{1}{\mathrm{Var}\, T}$.

Proof. The natural parametrization $\eta = c(\theta)$ gives

$$f^*(x \mid \eta) = \exp(\eta T(x) + d_0(\theta) + s(x))\mathbf{1}_A(x). \tag{2.31}$$

Log gives us

$$\log f^*(x \mid \eta) = \eta T(x) + d_0(\theta) + s(x) \implies \frac{\partial}{\partial\eta} \log f^*(x \mid \eta) = T(x) + d_0'(\eta) = T(x) - E_\eta(T). \tag{2.32}$$

Here $E_\eta(T) = -d_0'(\eta)$ and $\mathrm{Var}_\eta(T) = -d_0''(\eta)$. Therefore, we have

$$I^*(\eta) = E_\eta[\frac{\partial}{\partial\eta} \log f^*(X \mid \eta)] = E_\eta[T(X) - E_\eta(T)] = \mathrm{Var}_\eta(T). \tag{2.33}$$

Here $\delta(\theta) = E_\theta[T] = -d_0'(\eta) = h(\eta)$.

$$\frac{d\eta}{d\delta} = \left(\frac{d\delta}{d\eta}\right)^{-1} = (-d_0''(\eta))^{-1} = \frac{1}{\mathrm{Var}_\eta(T)} \implies I(\delta) = I^*(\eta)\left(\frac{d\eta}{d\delta}\right)^2 \Big|_{\eta=h^{-1}(\delta)} = \frac{1}{\mathrm{Var}_\eta(T)}. \tag{2.34}$$

■

2.2.1 Information Inequality

Suppose the conditions (A) and (B) (above) hold, and $0 < I(\theta) < \infty$. Let $T(X)$ be any statistic with $\mathrm{Var}\, T < \infty$ and such that the derivative, with respect to θ, of

$$E_\theta(T) = \int T(x)f(x \mid \theta)dx \tag{2.35}$$

exists and can be obtained by differentiating under the integral sign. Then

$$\mathrm{Var}_\theta(T(X)) \geq \frac{\frac{d}{d\theta}[E_\theta[T]]^2}{I(\theta)}. \tag{2.36}$$

This is known as the *Cramér-Rao lower bound*. To see this inequality, note that

$$\frac{d}{d\theta}E_\theta[T] = \int T(x)\frac{d}{d\theta}f(x \mid \theta)dx = \int T(x)S(x)f(x \mid \theta)dx = E_\theta[T(X)S(X)]. \tag{2.37}$$

$E[S(X)] = 0$, so the covariance $\mathrm{Cov}(T(X), S(X)) = E[T(X)S(X)] - E[T(X)]E[S(X)] = E[T(X)S(X)] \leq \sqrt{\mathrm{Var}(T(X))\,\mathrm{Var}(S(X))}$. $\mathrm{Var}(S(X))$ is nothing but $I(\theta)$. The inequality follows.

For the class of all unbiased estimators of θ we have

$$\mathrm{Var}_\theta(T(X)) \geq \frac{\frac{d}{d\theta}[E_\theta[T]]^2}{I(\theta)} = \frac{1}{I(\theta)}. \tag{2.38}$$

This lower bound is independent of any particular T. If there exists an unbiased estimator that attains this lower bound, then that estimator has to be the UMVUE.

Remark 2.23. Let $X_1, \ldots, X_n$ be i.i.d. from a location parameter family with probability distribution function $f_X(x \mid \theta) = f(x - \theta)$ with $\theta \in \mathbb{R}$. Then $R = X_{(n)} - X_{(1)}$ is ancillary.
If they come from a scale parameter family with pdf $f_X(x \mid \theta) = \frac{1}{\theta}f\left(\frac{x}{\theta}\right)$, then $R = X_{(n)} - X_{(1)}$ is also ancillary.

2.3 Order Statistics

Suppose $X_1, \ldots, X_n \sim F(x)$ are independent and identically distributed. Then the data points are arranged in increasing order and labelled $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. The distribution of the smallest order statistic $X_{(1)}$ is given by

$$F_{(1)}(x) = P(X_{(1)} \leq x) = 1 - P(X_{(1)} > x) = 1 - P(X_1 > x)^n = 1 - (1 - F(x))^n. \tag{2.39}$$

Similarly, the distribution of the largest order statistic $X_{(n)}$ is given by

$$F_{(n)}(x) = P(X_{(n)} \leq x) = P(X_1 \leq x)^n = (F(x))^n. \tag{2.40}$$

The joint distribution of $X_{(1)}$ and $X_{(n)}$ may also be obtained. Here,

$$\begin{aligned} P(X_{(n)} \leq y) &= P(X_{(1)} \leq x,\ X_{(n)} \leq y) + P(X_{(1)} > x,\ X_{(n)} \leq y) \\ \implies F_{X_{(1)},X_{(n)}}(x, y) &= P(X_{(1)} \leq x,\ X_{(n)} \leq y) = (F(y))^n - (F(y) - F(x))^n. \end{aligned} \tag{2.41}$$

Chapter 3

MULTIVARIATE NORMAL DISTRIBUTION

3.1 Random Vectors and Random Matrices

September 19th.

Definition 3.1. A *σ-algebra* (or *σ-field*) on a set Ω is a collection $\mathcal{F}$ of subsets of Ω that satisfies the following properties:

1. $\Omega \in \mathcal{F}$ (The entire set is in $\mathcal{F}$),
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (Closed under complementation),
3. If $A_1, A_2, A_3, \cdots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ (Closed under countable unions).

Let $(\Omega, \mathcal{F})$ be such a σ-algebra. One then defines a random variables across more than one dimension as follows:

Definition 3.2. The random variable X defined as $X = (X_1, \ldots, X_p) : \Omega \to \mathbb{R}^p$ for all $\omega \in \Omega$ is called a *p-variate random variable* if $X^{-1}(I_a) \in \mathcal{F}$ for all $I_a = \{(x_1, \ldots, x_p) \in \mathbb{R}^p \mid \infty < x_i \leq a_i, i = 1, \ldots, p\}$ and for all $a = (a_1, \ldots, a_p) \in \mathbb{R}^p$.
The *mean vector* is defined as $\mu = E[X] = (E[X_1], \ldots, E[X_p]) = (\mu_1, \ldots, \mu_p)$.

Definition 3.3. A random matrix $Z_{p\times q} = (Z_{ij})_{p\times q}$ is a $p \times q$ matrix of random variables, where each entry Z_{ij} is a random variable.
The *mean matrix* of Z is defined as $\mu_Z = E[Z] = (E[Z_{ij}])_{p\times q}$, where $E[Z_{ij}]$ is the expected value of the random variable Z_{ij} for all $i = 1, \ldots, p$ and $j = 1, \ldots, q$.

Note that if $G(Z)$ is a function of the random matrix Z, then the expectation is simply $E[G(Z)] = (E[G(Z)_{ij}])_{p\times q}$. A few more properties may be noted.

1. If $G(Z) = AZB$ where A and B are constant matrices of appropriate dimensions, then $E[G(Z)] = AE[Z]B$.
2. If (Z, T) has a joint distribution and A, B, C, and D are constant matrices of appropriate dimensions, then $E[AZB + CTD] = AE[Z]B + CE[T]D$.
3. If Z is symmetric and positive semidefinite with probability 1, then $E[Z]$ is also symmetric and positive semidefinite.

The first two properties follow from the linearity of the expectation in each of the entries of the matrix. The third property follows from the fact that if Z is positive semidefinite with probability 1, then for any vector x, we have $x^T Zx \geq 0$ with probability 1. Taking the expectation, we get $E[x^T Zx] = x^T E[Z]x \geq 0$, which implies that $E[Z]$ is also positive semidefinite. Similarly, one can show that $E[Z]$ is symmetric if Z is symmetric with probability 1.

Example 3.4. Suppose $Z_{p\times p}$ is symmetric and positive semidefinite with probability 1. Then its spectral decomposition gives gives $Z = \Gamma D_\lambda \Gamma^t$ where Γ is an orthogonal matrix and D_λ is a diagonal matrix with nonnegative entries. If $\lambda_i(Z)$ denotes the i^{th} diagonal entry of D_λ, then $\lambda_1(Z) \geq \lambda_2(Z) \geq \ldots \geq \lambda_p(Z) \geq 0$ with probability 1. However $\lambda_i(E[Z])$ need not be equal to $E[\lambda_i(Z)]$ for $i = 1, \ldots, p$. This is because the eigenvalues are not linear functions of the entries of the matrix. What is true is that $\lambda_i(E[Z]) \geq 0$.

3.1.1 Covariance Matrix

Definition 3.5. Suppose the random vector $X_{p\times 1}$ has mean μ, and

$$E[(X_i - \mu_i)(X_j - \mu_j)] = \text{Cov}(X_i, X_j) = \sigma_{ij} < \infty \tag{3.1}$$

for all $i, j = 1, \ldots, p$. Then the *covariance matrix* of X is defined as

$$\text{Cov}(X) = \Sigma = E[(X - \mu)(X - \mu)^t] = (E[(X_i - \mu_i)(X_j - \mu_j)])_{p\times p} = (\sigma_{ij})_{p\times p}. \tag{3.2}$$

One may also call the matrix the dispersion matrix or the variance-covariance matrix.

Theorem 3.6. *A matrix $\Sigma_{p\times p}$ is a covariance matrix of some random vector X if and only if Σ is symmetric and positive semidefinite.*

Proof. Suppose Σ is a covariance matrix of some random vector X. Then $\Sigma = E[(X - \mu)(X - \mu)^t]$ for some mean vector μ. Then for any $\alpha \in \mathbb{R}^p$, we have

$$\alpha^t \Sigma \alpha = \alpha^t E[(X - \mu)(X - \mu)^t]\alpha = E[\alpha^t (X - \mu)(X - \mu)^t \alpha] = E[(\alpha^t (X - \mu))^2] = \text{Var}(\alpha^t X) \geq 0 \tag{3.3}$$

showing positive semidefinitness. Also, $\Sigma^t = (E[(X - \mu)(X - \mu)^t])^t = E[(X - \mu)(X - \mu)^t] = \Sigma$ showing symmetry. For the converse, suppose Σ is symmetric and positive semidefinite. Also suppose it has rank $r \leq p$. Then $\Sigma = CC^t$ for some $C_{p\times r}$ with rank r. Let $Y_1, \ldots, Y_r$ be independent and identically distribution $N(0, 1)$ random variables, and let $X = CY$. Then $E[X] = E[CY] = CE[Y] = 0$ and

$$\text{Cov}(X) = E[XX^t] = E[CY(CY)^t] = CE[YY^t]C^t = CI_r C^t = CC^t = \Sigma \tag{3.4}$$

where $E[YY^t] = \text{Cov}(Y) = I_r$ since the Y_i are independent $N(0, 1)$ random variables. Thus Σ is a covariance matrix of the random vector X. ∎

Let $X_{p\times 1}$ and $Y_{q\times 1}$ be jointly distributed with finite second moments for their elements and with $E[X] = \mu$ and $E[Y] = \nu$. Then

1. $\text{Cov}(X, Y) = E[(X - \mu)(Y - \nu)^t]$ is the $p \times q$ covariance matrix between X and Y.
2. $\text{Cov}(X) = E[XX^t] = E[X]E[X]^t$.
3. $\text{Cov}(AX, BY) = A\,\text{Cov}(X, Y)B^t$ for any constant matrices A and B of appropriate dimensions.
4. $\text{Cov}(AX) = A\,\text{Cov}(X)A^t$ for any constant matrix A of appropriate dimensions.

The *moment generating function* of a random vector $X_{p\times 1}$ is defined, simply, as

$$\phi_X(\alpha) = E[e^{\alpha^t X}] \text{ for all } \alpha \in \mathbb{R}^p. \tag{3.5}$$

If X_1 and X_2 are independent, then

$$\phi_{X_1+X_2}(\alpha) = E[e^{\alpha^t (X_1+X_2)}] = E[e^{\alpha^t X_1} e^{\alpha^t X_2}] = E[e^{\alpha^t X_1}]E[e^{\alpha^t X_2}] = \phi_{X_1}(\alpha)\phi_{X_2}(\alpha). \tag{3.6}$$

Theorem 3.7 (*Cramér-Wold device*). *If $X_{p\times 1}$ is a random vector, then its probability distribution is completely determined by the distribution of all linear functions $\alpha^t X$ for all $\alpha \in \mathbb{R}^p$.*

3.2 Multivariate Normal Distribution

Definition 3.8. A random vector $X_{p\times 1}$ is said to be *p-variate normally distributed* if for every $\alpha \in \mathbb{R}^p$, the distribution of $\alpha^t X$ is univariate normal.

If X has the p-variate normal distribution then both $\mu = E[X]$ and $\Sigma = \text{Cov}(X)$ exist and are finite. We claim that the distribution of X is completely determined by μ and Σ. TO see this, let $X = (X_1, \ldots, X_p)^t$. Then $e_i^t X = X_i$ is univariate normal for each $i = 1, \ldots, p$ where e_i is the i^{th} standard basis vector in $\mathbb{R}^p$. Thus each X_i is univariate normal and follows distribution $N(\mu_i, \sigma_{ii}^2)$. Set $\mu = (\mu 1, \ldots, \mu_p)^t$ and $\Sigma = (\sigma_{ij})_{p\times p}$. Furhter, $E[\alpha^t X] = \alpha^t \mu$ and $\text{Var}(\alpha^t X) = \alpha^t \Sigma \alpha$. Thus $\{\alpha^t X \mid \alpha \in \mathbb{R}^p\}$ determines the distribution of X by the Cramér-Wold device.

IF $X = N_p(\mu, \Sigma)$, then for any $A_{k\times p}$ and $b_{k\times 1}$, we have

$$Y = AX + b \sim N_k(A\mu + b, A\Sigma A^t). \tag{3.7}$$

Theorem 3.9. *$X_{p\times 1} \sim N_p(\mu, \Sigma)$ if and only if $X_{p\times 1} = C_{p\times r} Z_{r\times 1} + \mu_{p\times 1}$ where $Z = (Z_1, \ldots, Z_r)^t$ with Z_i's following independent $N(0,1)$ distributions, $C_{p\times r}$ is a constant matrix with rank $r \le p$, and $\Sigma = CC^t$.*

Proof. Suppose $Z \sim N_r(0, I_r)$ and $X = CZ + \mu$ with $C_{p\times r}$ a constant matrix with rank $r \le p$, and $\mu \in R^p$. The characteristic function of X is

$$\phi_X(t) = E[e^{it^T X}] = E[e^{it^T X}] = E[e^{it^T (CZ+\mu)}] = e^{it^T \mu} E[e^{i(C^T t)^T Z}] = e^{it^T \mu} e^{-\frac{1}{2} t^T CC^T t}. \tag{3.8}$$

■

Theorem 3.10. *If $X \sim N_p(\mu, \Sigma)$, then the marginal distribution of any subset of k components of X is k-variate normal.*

Proof. Suppose, without the loss of generality, that the first k components of X are considered. Then we can write $X = (X^{(1)}_{k\times 1}, X^{(2)}_{(p-k)\times 1})^t$ where $X^{(1)} = (X_1, \ldots, X_k)^t$ and $X^{(2)} = (X_{k+1}, \ldots, X_p)^t$. Then we can write $X^{(1)} = (I_{k\times k}, 0_{k\times(p-k)})X$. Thus $X^{(1)}$ is a linear function of X and hence is k-variate normal. ■

Proposition 3.11. *Suppose*

$$X_{p\times 1} = \begin{pmatrix} X^{(1)}_{k\times 1} \\ X^{(2)}_{(p-k)\times 1} \end{pmatrix} \sim N_p \left(\begin{pmatrix} \mu^{(1)}_{k\times 1} \\ \mu^{(2)}_{(p-k)\times 1} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right). \tag{3.9}$$

Then $X^{(1)}$ and $X^{(2)}$ are independent if and only if $\Sigma_{12} = 0$ (or equivalently $\Sigma_{21} = 0$).

Proof. To prove the result, we use the moment generating function (MGF) of $X = (X^{(1)}, X^{(2)})^T$. The joint distribution of $X^{(1)}$ and $X^{(2)}$ is multivariate normal, so the MGF of X is given by:

$$M_{(X^{(1)}, X^{(2)})}(s_1, s_2) = E[e^{s_1^T X^{(1)} + s_2^T X^{(2)}}]. \tag{3.10}$$

This can be written as:

$$M_{(X^{(1)}, X^{(2)})}(s_1, s_2) = E[e^{(s_1^T, s_2^T)^T X}] = e^{(s_1^T, s_2^T)^T \mu + \frac{1}{2}(s_1^T, s_2^T)^T \Sigma (s_1^T, s_2^T)}. \tag{3.11}$$

Expanding the quadratic form:

$$(s_1^T, s_2^T)^T \Sigma (s_1^T, s_2^T) = s_1^T \Sigma_{11} s_1 + s_1^T \Sigma_{12} s_2 + s_2^T \Sigma_{21} s_1 + s_2^T \Sigma_{22} s_2. \tag{3.12}$$

So the MGF becomes:

$$M_{(X^{(1)},X^{(2)})}(s_1,s_2) = e^{s_1^T\mu^{(1)}+s_2^T\mu^{(2)}+\frac{1}{2}\left(s_1^T\Sigma_{11}s_1+s_1^T\Sigma_{12}s_2+s_2^T\Sigma_{21}s_1+s_2^T\Sigma_{22}s_2\right)}. \tag{3.13}$$

Now, recall that $X^{(1)}$ and $X^{(2)}$ are independent if and only if their joint distribution factors, i.e., the cross terms involving both $X^{(1)}$ and $X^{(2)}$ vanish. For the MGF to factorize into a product of two functions (one depending only on s_1 and the other only on s_2), the cross terms must disappear. These cross terms are given by:

$$s_1^T\Sigma_{12}s_2 + s_2^T\Sigma_{21}s_1. \tag{3.14}$$

Since Σ_{12} is a $k \times (p-k)$ matrix and $\Sigma_{21} = \Sigma_{12}^T$, we have:

$$s_1^T\Sigma_{12}s_2 + s_2^T\Sigma_{21}s_1 = 2s_1^T\Sigma_{12}s_2. \tag{3.15}$$

For the MGF to factor, we need this term to vanish for all s_1 and s_2. This occurs if and only if:

$$\Sigma_{12} = 0. \tag{3.16}$$

Therefore, $X^{(1)}$ and $X^{(2)}$ are independent if and only if $\Sigma_{12} = 0$. Since $\Sigma_{12} = \Sigma_{21}^T$, this condition is equivalent to $\Sigma_{21} = 0$. ■

Theorem 3.12. *Suppose $X \sim N_p(\mu,\Sigma)$ with Σ positive definite. Then the probability density function of X is given by*

$$f_X(x) = \frac{1}{(2\pi)^{\frac{p}{2}}}\,|\Sigma|^{-\frac{1}{2}}\,e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)} \text{ for all } x \in \mathbb{R}^p. \tag{3.17}$$

Proof. Let $\Sigma = CC^t$ with $C = \Sigma^{\frac{1}{2}}$ is non-singular. Then $X = CZ + \mu$ where $Z \sim N_r(0, I_r)$. Thus,

$$f_Z(z) = (2\pi)^{-\frac{r}{2}}e^{-\frac{1}{2}z^tz} \implies f_X(x) = f_Z(C^{-1}(x-\mu))\left|\det(C^{-1})\right| = \frac{(2\pi)^{-\frac{r}{2}}}{|\det(C)|}e^{-\frac{1}{2}(x-\mu)^t(CC^t)^{-1}(x-\mu)}. \tag{3.18}$$

■

Theorem 3.13. *Let $X \sim N_p(\mu,\Sigma)$ with Σ positive definite, and let $X = (X_1,X_2)^t$, $\mu = (\mu_1,\mu_2)^t$, and $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ where X_1 and μ_1 are of length k. Let $\Sigma_{11\cdot2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.*

1. *If $\Sigma_{11\cdot2}$ is positive definite, then $X_1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \sim N_k(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\mu_2, \Sigma_{11\cdot2})$ and is independent of X_2.*

Proof. 1. Let $X \sim N_p(\mu,\Sigma)$, $A \in \mathbb{R}^{m\times p}$ and $b \in \mathbb{R}^m$. Then $AX + b \sim N_m(A\mu + b, A\Sigma A^t)$. Set $C = \begin{pmatrix}I_k & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I_{p-k}\end{pmatrix}$. Then

$$CX = \begin{pmatrix}I_k & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I_{p-k}\end{pmatrix}\begin{pmatrix}X_1\\X_2\end{pmatrix} = \begin{pmatrix}X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2\\ X_2\end{pmatrix} \tag{3.19}$$

■

Chapter 4

CONVERGENCE

4.1 Types of Convergence

September 26th.

Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and let X be another random variable defined on the same probability space.

Definition 4.1. Notion of *convergence in probability*: We say that X_n converges to X, $X_n \xrightarrow{P} X$, if for every $\varepsilon > 0$,

$$P(|X_n - X| \geq \varepsilon) \to 0 \text{ as } n \to \infty. \tag{4.1}$$

An example is the *weak law of large numbers* which states that if $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean μ, then the sample average $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ.

Definition 4.2. Notion of *almost sure convergence*: We say that X_n converges to X almost surely, $X_n \xrightarrow{a.s.} X$, if

$$P\left(\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1. \tag{4.2}$$

An example is the *strong law of large numbers* which states that if $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean μ, then the sample average $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to μ.

Definition 4.3. Notion of *convergence in distribution*: We say that X_n converges to X in distribution, $X_n \xrightarrow{d} X$, if

$$F_{X_n}(x) \to F_X(x) \text{ at all points } x \text{ where } F_X \text{ is continuous,} \tag{4.3}$$

where F_{X_n} and F_X are the cumulative distribution functions of X_n and X respectively.

An example is the *central limit theorem* which states that if $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean μ and variance σ^2, then the standardized sum $\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}}$ converges in distribution to a standard normal random variable as $n \to \infty$.

Lemma 4.4 (*Borel-Cantelli lemma*)**.** *Let $A_1, A_2, \ldots$ be a sequence of events in a probability space $(\Omega, \mathcal{F}, P)$.*

1. *If the sum of the probability of the events A_n is finite, i.e., $\sum_{n=1}^{\infty} P(A_n) < \infty$, then the probability that infinitely many of the events A_n occur is zero, i.e., $P(\limsup_{n\to\infty} A_n) = 0$.*
2. *If $\sum_{n=1}^{\infty} P(A_n) = \infty$ and the events A_n are independent, then the probability that infinitely many of the events A_n occur is one, i.e., $P(\limsup_{n\to\infty} A_n) = 1$.*

Theorem 4.5 (*Slutsky's theorem*)**.** *Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, where c is a constant. Then*

1. $X_n + Y_n \xrightarrow{d} X + c$,
2. $X_n Y_n \xrightarrow{d} Xc$, *and*
3. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$, *provided* $c \neq 0$.

Proof. With the conditions, one can show that the joint vector (X_n, Y_n) converges in distribution to (X, c). The results then follow from the continuous mapping theorem, which states that for a continuous function g, $Z_n \xrightarrow{d} Z$ implies $g(Z_n) \xrightarrow{d} g(Z)$. ■

Corollary 4.6. *If $X_n - Y_n \xrightarrow{P} 0$ and $X_n \xrightarrow{d} X$, then $Y_n \xrightarrow{d} X$.*

Proof. Work similarly as before, looking at the joint vector $(X_n - Y_n, X_n)$ and using the continuous mapping theorem with the function $g(x, y) = y - x$. ■

Recall *Markov's inequality* where for any random variable X and $a > 0$, we have

$$P(X \geq a) \leq \frac{EX}{a}. \tag{4.4}$$

Replacing X by X^2 and a with a^2 gives

$$P(|X| \geq a) \leq \frac{E[X^2]}{a^2}. \tag{4.5}$$

Finally, replacing X by $|X - E[X]|$ gives *Chebyshev's inequality*:

$$P(|X - EX| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}. \tag{4.6}$$

Theorem 4.7 (*Kinchin's weak law of large numbers*)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and indentically distributed random variables such that $EX_i = \mu$ exists and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\overline{X}_n \xrightarrow{P} \mu$.*

Theorem 4.8 (*Chebyshev's weak law of large numbers*)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables with $EX_i = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2 < \infty$. Moreover, suppose $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \to 0 \text{ as } n \to \infty \implies \overline{X}_n - \overline{\mu}_n \xrightarrow{P} 0, \tag{4.7}$$

where $\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i$.

Theorem 4.9 (*Kolmogorov's strong law of large numbers*)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables such that $EX_i = \mu$ exists. Then $\overline{X}_n \xrightarrow{a.s.} \mu$.*

Theorem 4.10 (*Lindeberg-levy central limit theorem*)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with $EX_i = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then*

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1). \tag{4.8}$$

4.1.1 Consistency

Let $X_1, X_2, \ldots$ be independent and identically distributed P_θ random variables with $\theta \in \Theta \subseteq \mathbb{R}$. An estimator $T_n(X_1, \ldots, X_n)$ of $q(\theta)$ is said to be a *consistent estimator* if

$$T_n \xrightarrow{P} q(\theta) \text{ for all } \theta \in \Theta. \tag{4.9}$$

Theorem 4.11. *If $\{T_n\}_{n\geq 1}$ is a sequence of estimators such that $E_\theta[T_n] \to q(\theta)$ and $\mathrm{Var}_\theta(T_n) \to 0$ as $n \to \infty$ for all $\theta \in \Theta$, then T_n is a consistent estimator of $q(\theta)$.*

Theorem 4.12 (The *invariance theorem*)**.** *If T_n is consistent for θ and h is a continuous function, then $h(T_n)$ is consistent for $h(\theta)$.*

It also follows that if $a_n \to 1$ and $b_n \to 0$, then $a_n T_n + b_n$ is consistent for θ for when T_n is consistent for θ. A few more properties may be noted:

1. An unbiased estimator may not be consistent. A counterexample is X_1 as an estimator of μ where $X_1, X_2, \ldots$ are independent and identically distributed $N(\mu, 1)$ random variables. Clearly, $E[X_1] = \mu$, but X_1 is not consistent.
2. A method of moments estimator is always consistent if the first moment exists.
3. A maximum likelihood estimator may not be consistent.
4. A consistent estimator may not be unique.

Example 4.13. Suppose $X_1, X_2, \ldots$ are independent and identically distributed $U[0, \theta]$ random variables. Then the estimator $T_n = 2\overline{X}_n$ is consistent for θ since the first moment is $\theta/2$ and $E[\overline{X}] = \theta/2$; via method of moments, the result follows.

Proposition 4.14. *Suppose $\{a_n\}_{n\geq 1}$ diverges to infinity and b is a constant. Also suppose that $a_n(X_n - b) \xrightarrow{d} X$ and g be a continuously differentiable function such that $g'(b)$ is non-zero. Then*

$$a_n(g(X_n) - g(b)) \xrightarrow{d} g'(b)X. \tag{4.10}$$

Proof. $X_n - b = \frac{1}{a_n} a_n(X_n - b) \xrightarrow{P} 0$ since $a_n \to \infty$ and $a_n(X_n - b) \xrightarrow{d} X$. By Slutsky's theorem, $X_n \xrightarrow{d} b$. Now there exists a random variable $X_n < X_n^* < b$ such that

$$a_n(g(X_n) - g(b)) = g'(X_n^*)(X_n - b), \tag{4.11}$$

by the mean value theorem. Since $X_n^* \xrightarrow{P} b$, $g'(X_n^*) \xrightarrow{P} g'(b)$. ■

4.1.2 Asymptotic Notion and Efficiency

Definition 4.15. Suppose $T_n(X_1, \ldots, X_n)$ is an unbiased estimator of $q(\theta)$ for all $\theta \in \Theta$. Then T_n is said to be *asymptotically normal* if

$$\sqrt{n}(T_n - q(\theta)) \xrightarrow{d} N(0, \sigma^2(\theta)) \text{ for all } \theta \in \Theta. \tag{4.12}$$

A consistent sequence of estimators is said to be *consistent and asymptotically normal* if

$$T_n \sim AN(q(\theta), \frac{v(\theta)}{n}) \text{ for all } \theta \in \Theta, \tag{4.13}$$

where $v(\theta) = \frac{1}{I(\theta)}$ is the inverse of the Fisher information matrix. T_n is termed the *best asymptotically normal estimator* if $v(\theta)$ is the minimum variance among all consistent and asymptotically normal estimators.

Definition 4.16. A sequence of estimators $T_n(X_1, \ldots, X_n)$ is said to be *asymptotically unbiased* if $E_\theta[T_n] \to q(\theta)$.

Definition 4.17. Let T_1, T_2 be two unbiased estimators for a parameter θ. Suppose that $E[T_1^2], E[T_2^2] < \infty$. The *relative efficiency* of T_1 relative to T_2 is defined as

$$\text{eff}_\theta(T_1 \mid T_2) = \frac{\operatorname{Var} T_2}{\operatorname{Var} T_1} \tag{4.14}$$

and we call T_1 more efficient than T_2 if $\text{eff}_\theta(T_1 \mid T_2) > 1$ for all $\theta \in \Theta$.

Definition 4.18. Assume that the regularity conditions of the FRC inequality are satisfied by the family of distributions $\{F_\theta \mid \theta \in \Theta\}$. We say that an unbiased estimator T for the parameter θ is *most efficient* if

$$\operatorname{Var} T = \left(E_\theta \left[\frac{\partial \log f_\theta(x)}{\partial \theta}^2 \right] \right)^{-1} = \frac{1}{I_\eta(\theta)}. \tag{4.15}$$

Definition 4.19. Let T be the most efficient estimator for the regular family of distributions $\{F_\theta \mid \theta \in \Theta\}$. Then the *efficiency* of an unbiased estimator T_1 of θ is defined as

$$\text{eff}_\theta(T_1) = \text{eff}_\theta(T_1 \mid T) = \frac{\operatorname{Var} T}{\operatorname{Var} T_1} = \frac{I_n^{-1}(\theta)}{\operatorname{Var} T_1}. \tag{4.16}$$

If $\lim_{n\to\infty} \text{eff}_\theta(T_n) = 1$, then T_n is said to be *asymptotically efficient*.

4.2 Asymptotically Normal

October 10th.

In the notion of asymptotic normality, if we take $T_n(\underline{X}) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, and $q(\theta) = \mu$ and $\sigma^2(\theta) = \sigma^2$, then we have

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2), \tag{4.17}$$

which is the central limit theorem. The delta method says that if we have a sequence of random variables X_n such that $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, and h is continuous and differentiable at μ with $h'(\mu) \neq 0$, then

$$\sqrt{n}(h(X_n) - h(\mu)) \xrightarrow{d} N(0, h'(\mu)^2\sigma^2). \tag{4.18}$$

Example 4.20. Let $X_1, X_2, \ldots$ be independent and identically distributed $\text{Exp}(\theta)$ random variables, $\theta > 0$. We wish to check $\hat{\theta} = \frac{1}{\overline{X}_n}$ is asymptotically normal. We have $EX = \frac{1}{\theta}$, and $\operatorname{Var} X = \frac{1}{\theta^2}$.

Since $\overline{X}_n \xrightarrow{P} \frac{1}{\theta}$, we can apply the central limit theorem to get

$$\sqrt{n}(\overline{X}_n - \frac{1}{\theta}) \xrightarrow{d} N(0, \frac{1}{\theta^2}). \tag{4.19}$$

We can apply the delta method via the function $h(x) = \frac{1}{x}$, which is continuous at $\frac{1}{\theta}$. Thus,

$$\sqrt{n}(\frac{1}{\overline{X}} - \theta) \xrightarrow{d} N(0, \theta^2). \tag{4.20}$$

$T_n = T_n(\underline{X})$ is said to be asympotically efficient for estimating $q(\theta)$ if its asymptotic variance is

$$\sigma^2(\theta) = \frac{(q'(\theta))^2}{I_n(\theta)}. \tag{4.21}$$

Definition 4.21. Suppose $T_n^{(1)}$ and $T_n^{(2)}$ are two estimators of $q(\theta)$ such that

$$\sqrt{n}(T_n^{(1)} - q(\theta)) \xrightarrow{d} N(0, \sigma_1^2(\theta)), \tag{4.22}$$

$$\sqrt{n}(T_n^{(2)} - q(\theta)) \xrightarrow{d} N(0, \sigma_2^2(\theta)). \tag{4.23}$$

Then the *asymptotic relative efficiency* of $T_n^{(1)}$ with respect to $T_n^{(2)}$ is define to be

$$e_\theta(T_n^{(1)}, T_n^{(2)}) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}. \tag{4.24}$$

$T_n^{(1)}$ is better if $\sigma_2^2 > \sigma_1^2$.

Example 4.22. Let $X_1, X_2, \ldots$ be independent and identically distributed $N(0, \sigma^2)$ random variables. Consider two estimators for σ:

$$\hat{\sigma}_1 = \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}, \quad \hat{\sigma}_2 = \sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^n |X_i|. \tag{4.25}$$

Note that $X_i^2 \sim \sigma^2\chi_1^2$, so $E[X_i^2] = \sigma^2$ and $\mathrm{Var}(X_i^2) = 2\sigma^4$. By the central limit theorem,

$$\sqrt{n}(\hat{\sigma}_1^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4). \tag{4.26}$$

Via the function $h(x) = \sqrt{x}$, we have $h'(\sigma^2) = \frac{1}{2\sigma}$, the delta method gives

$$\sqrt{n}(\hat{\sigma}_1 - \sigma) \xrightarrow{d} N(0, \frac{\sigma^2}{2}). \tag{4.27}$$

Since $Z_i = \frac{X_i}{\sigma} \sim N(0,1)$, $E[|Z_i|] = \sqrt{\frac{2}{\pi}}$ and $\mathrm{Var}(|Z_i|) = 1 - \frac{2}{\pi}$. Thus, $E[|X_i|] = \sigma\sqrt{\frac{2}{\pi}}$ and $\mathrm{Var}(|X_i|) = \sigma^2(1 - \frac{2}{\pi})$. By the central limit theorem,

$$\sqrt{n}(\hat{\sigma}_2^2 - \sigma^2) \xrightarrow{d} N(0, \sigma^2(\frac{\pi}{2} - 1)). \tag{4.28}$$

Delta method via $h(x) = \sqrt{x}$ gives

$$\sqrt{n}(\hat{\sigma}_2 - \sigma) \xrightarrow{d} N(0, (\frac{\pi}{2} - 1)\sigma^2). \tag{4.29}$$

Chapter 5

HYPOTHESIS TESTING AND INTERVAL ESTIMATION

5.1 Hypothesis Testing

Recall that given a null hypothesis H_0, then the *significance level* α is the probability of a Type I error; that is,

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true}). \tag{5.1}$$

β is defiend as the probability of a Type II error; that is,

$$\beta = P(\text{accept } H_0 \mid H_0 \text{ is false}). \tag{5.2}$$

We start with an example.

Example 5.1. A pack of a certain brand of cigarettes displays the statement, '1.5 mg nicotine on average per cigarette'. Let μ be the actual nictone content per cigarette for this brand. It is required to test if the actual average is higher than what is claimed. Suppose a sample of cigarettes is selected and the nicotine content per cigarette is measured. Let $X_1, X_2, \ldots, X_n$ be the nicotine content of the n cigarettes in the sample. We assume that $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. We consider the null hypothesis $H_0 : \mu = 1.5$ and the alternate $H_1 : \mu > 1.5$. In this case, we consider the test statistic

$$T = \frac{\overline{X} - 1.5}{s/\sqrt{n}} \tag{5.3}$$

If H_0 is true, then $T \sim t_{n-1}$. We reject H_0 if $T > t_{n-1,\alpha}$, where $t_{n-1,\alpha}$ is the $100(1-\alpha)$ percentile of the t-distribution with $n-1$ degrees of freedom.

5.1.1 Neyman Pearson Theory of Testing

Let $X \sim P_\theta$, $\theta \in \Theta$. Let $\mathcal{X}$ be the sample space of X, that is, $\mathcal{X} = X(\Omega)$. We wish to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where $\Theta_0, \Theta_1 \subseteq \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. A simple hypothesis is one where Θ_0 is a singleton.

In a non randomized test, find a subset of S of $\mathcal{X}$ and reject H_0 of the observed value $x \in S$. $S \subset \mathcal{X}$ is called the critical region or the rejection region. The test function ϕ for randomized tests is defined as $\phi(x) = \mathbf{1}_S(x)$. For a level of α test, we require that

$$\sup_{\theta \in \Theta_0} P_\theta(X \in S) \leq \alpha. \tag{5.4}$$

If $\Theta_0 = \{\theta_0\}$, then we simply require $P_{\theta_0}(X \in S) \leq \alpha$. For a randomized test, any ϕ with $0 \leq \phi(x) \leq 1$ for all $x \in \mathcal{X}$ and at any x, $\phi(x)$ is the probability of rejecting H_0 when $X = x$. A non randomized test

is a subset of a randomized test. The power function of the test is defined as

$$P_\theta(\text{reject} H_0) = E_\theta(P(\text{reject } H_0 \mid X)) = E_\theta[\phi(X)] = \int_{\mathcal{X}} \phi(x) f(x \mid \theta) dx. \tag{5.5}$$

The problem is to find ϕ such that $E_\theta[\phi(X)]$ is maximized when $\theta \in \Theta_1$ subject to the condition that $\sup_{\theta \in \Theta_0} E_\theta[\phi(X)] \leq \alpha$. Such a test, if it exists, is called the *uniformly most powerful* (UMP) test of level α.

Lemma 5.2. Neyman-Pearson lemma. *Suppose $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Let P_{θ_1} and P_{θ_2} be the respective densities. Then*

1. *there exists a test ϕ and a constant $\alpha \geq 0$ such that $E_\theta[\phi(X)] = \alpha$.*
2. *If a test satisfies the above and below:*

$$\phi(x) = \begin{cases} 1 & \text{if } P_{\theta_1}(x) > kP_{\theta_0}(x), \\ 0 & \text{if } P_{\theta_1}(x) < kP_{\theta_0}(x). \end{cases} \tag{5.6}$$

   for some $k \geq 0$, then it is the most powerful for testing $H_0 : P_\theta = P_{\theta_0}$ against $H_1 : P_\theta = P_{\theta_1}$ at level α.
3. *Conversely, if ϕ is a most powerful test at level α for testing, then it satisfies the above for some $k \geq 0$.*

Proof. Pg no. 80 of Mohan Delampady's notes. ■

5.1.2 Likelihood Ratio Test

Definition 5.3. P_θ, $\theta \in \Theta \subseteq \mathbb{R}$ with density $f(x \mid \theta)$ is said to have *monotone likelihood ratio* (MLR) if there exists a real valued function $T(x)$ such that for any $\theta < \theta'$, $P_\theta \neq P_{\theta'}$, and the likelihood ratio $\frac{f(x|\theta')}{f(x|\theta)}$ is a non decreasing function of $T(x)$, that is,

$$\frac{f(x \mid \theta')}{f(x \mid \theta)} = h_{\theta,\theta'}(T(x)), \text{ where } h_{\theta,\theta'} \text{ is non decreasing.} \tag{5.7}$$

October 17th.

Theorem 5.4. *Let θ be a real parameter and let X have density $f(x \mid \theta)$ with MLR in $T(x)$. Then*

1. *for every $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$, there exists a UMP test of level α given by*

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > c, \\ \gamma & \text{if } T(x) = c, \\ 0 & \text{if } T(x) < c, \end{cases} \tag{5.8}$$

   where c and γ are determined by $E_{\theta_0}[\phi(X)] = \alpha$.
2. *The power function $E_\theta[\phi(X)]$ of this test is strictly increasing for all θ satisfying $E_\theta[\phi(X)] < 1$.*
3. *For all θ', the test given by part 1. is the UMP for testing $H_0 : \theta \leq \theta'$ and $H_1 : \theta > \theta'$ at level $\alpha' = E_{\theta'}[\phi(X)]$.*
4. *For any $\theta < \theta_0$, the test given by part 1. minimizes $E_\theta(\phi(X))$ among all test satisfying $E_{\theta_0}[\phi(X)] = \alpha$.*

Example 5.5. We have $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known. We wish to test $H_0 : \mu \leq \mu_0$

against $H_1 : \mu > \mu_0$. With $T(X) = \overline{X}$, we have MLR. Thus, the UMP test of level α is given by

$$\phi(X) = \begin{cases} 1 & \text{if } \overline{X} > c, \\ \gamma & \text{if } \overline{X} = c, \\ 0 & \text{if } \overline{X} < c, \end{cases} \tag{5.9}$$

where γ does not matter since $P(\overline{X} = c) = 0$. To find c, we require that $E_{\mu_0}[\phi(X)] = \alpha$ when $\mu = \mu_0$. Thus,

$$\alpha = E_\theta[\phi(X)] = P(\overline{X} > c) = P\left(Z > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = z_{1-\alpha}. \tag{5.10}$$

Thus, $c = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. We now wish to test $H_0 : \sigma^2 \leq \sigma_0^2$ against $H_1 : \sigma^2 > \sigma_0^2$. With $T(X) = \sum_{i=1}^n (X_i - \mu_0)^2$, we have MLR. Thus, the UMP test of level α is given by $\phi(X) = \mathbf{1}_{\{T(X)>c\}}$, where c is determined by a chi-squared distribution.

5.1.3 Generalized Likelihood Ratio Test

Let $X \sim P_\theta$, $\theta \in \Theta$, have ddensity $f(x \mid \theta)$. Consider the test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where $\Theta_0, \Theta_1 \subseteq \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The *generalized likelihood ratio* (GLR) is defined as

$$L(x) = \frac{\sup_{\theta\in\Theta_1} f(x \mid \theta)}{\sup_{\theta\in\Theta_0} f(x \mid \theta)}. \tag{5.11}$$

Consider

$$\lambda(x) = \frac{\sup_{\theta\in\Theta} f(x \mid \theta)}{\sup_{\theta\in\Theta_0} f(x \mid \theta)}. \tag{5.12}$$

Note that $\lambda(x) = \max\{1, L(x)\}$. The GLR test rejects H_0 if $L(x) > k$ for some $k \geq 0$, or equivalently, if $\lambda(x) > c$ for some $c \geq 1$.

Example 5.6. $X_i \sim N(\mu, \sigma^2)$, with μ and σ^2 unknown. We wish to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. We have $\Theta_0 = \{(0, \sigma^2) : \sigma^2 > 0\}$ and $\Theta_1 = \{(\mu, \sigma^2) : \mu \neq 0, \sigma^2 > 0\}$. The MLE are needed to compute the GLR statistic. We have

$$\lambda(\underline{x}) = \frac{f(x \mid \hat{\theta}_1)}{f(x \mid \hat{\theta}_0)} = \frac{(2\pi)^{-n/2}(-\frac{1}{n}\sum(x_i - \overline{x})^2)^{-n/2}\exp(-\frac{1}{2\hat{\sigma_1}^2}(\sum(x_i - \overline{x}^2) + n(\overline{x} - \hat{\mu}_1^2)))}{(2\pi)^{-n/2}(-\frac{1}{n}\sum(x_i - \overline{x})^2)^{-n/2}\exp(-\frac{1}{2\hat{\sigma_0^2}}(\sum(x_i - \overline{x})^2 + n(\overline{x} - 0)^2))}. \tag{5.13}$$

5.1.4 Confidence Set and Interval Estimation

For a *confidence set*, we want $S(X) \subseteq \Theta$ such that $P(\theta \in S(X)) \geq 1 - \alpha$ for all $\theta \in \Theta$. $S(X)$ is then said to be a $100(1-\alpha)\%$ confidence set for θ. Here we test $H_0 : \theta = \theta'$ against $H_1 : \theta \neq \theta'$ for any $\theta' \in \Theta$. Let $A(\theta') \subseteq \mathcal{X}$ be the acceptance region of level α test for this hypothesis. Define $S(X) = \{\theta' \in \Theta \mid X \in A(\theta')\}$.

Example 5.7. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known. $\overline{X}$ here is sufficient for μ. We wish to find a $100(1-\alpha)\%$ confidence interval for μ. For any $\mu' \in \mathbb{R}$, we test $H_0 : \mu = \mu'$ against $H_1 : \mu \neq \mu'$. The level α test has acceptance region $A(\mu') = \{x : |\overline{x} - \mu'| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}$. Thus the confidence set (interval in our case) is

$$S(\overline{X}) = \left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{5.14}$$

Let $X \sim P_\theta$, $\theta \in \Theta$.

Definition 5.8. An interval $(\underline{T}(x), \overline{T}(x))$, with $\underline{T}(x) < \overline{T}(x)$ for all $x \in \mathcal{X}$, is said to be a *confidence interval* for $q(\theta)$ if

$$\inf_{\theta \in \Theta} P_\theta(\underline{T}(X) < q(\theta) < \overline{T}(X)) \geq 1 - \alpha. \tag{5.15}$$

A closed interval may also be termed so.

To define confidence intervals, we use the concept of *pivotal quantity*. It is simply a real valued function $T(\underline{X}, \theta)$ such that the distribution of $T(\underline{X}, \theta)$ does not depend on θ. Suppose we choose two numbers t_1 and t_2 such that $P_\theta(t_1 \leq T(\underline{x}, \theta) \leq t_2) = 1 - \alpha$. If, for each $\underline{x}$, $T(\underline{x}, \theta)$ is monotone in θ, then we can solve the inequalities $t_1 \leq T(\underline{x}, \theta) \leq t_2$ for θ to get the confidence interval.

For example, if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, with both parameters unknown, then $T(\underline{X}) = \frac{\sum (X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$ is a pivotal quantity for σ^2. Choosing t_1 and t_2 such that $P(t_1 \leq \chi^2_{n-1} \leq t_2) = 1 - \alpha$, we have as desired.

5.2 Bayesian Inference

October 24th.

Model: $X \mid \theta$ has density $f(x \mid \theta)$, $\theta \in \Theta$. The *prior*: θ has density $\pi(\theta)$. The *posterior*: the density is

$$\pi(\theta \mid \underline{x}) = \frac{f(\underline{x} \mid \theta)\pi(\theta)}{m(\underline{x})} \tag{5.16}$$

where $m(\underline{x}) = \int_\Theta f(\underline{x} \mid \theta)\pi(\theta) d\theta$ is the marginal density of $\underline{x}$. Given all the ingredients, the Bayesian calculates the conditional probability density of θ given $\underline{X} = \underline{x}$. The joint density of x and θ is

$$h(x, \theta) = f(x \mid \theta)\pi(\theta). \tag{5.17}$$

Theorem 5.9. *Suppose $T = T(\underline{X})$ is sufficient for θ. Then the posterior distribution of θ given $X = x$ depends on x only through $T(x)$.*

We have $f(x \mid \theta) = g(T(x), \theta)h(x)$, thus for $T(x) = t$,

$$\pi(\theta \mid \underline{x}) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid u)\pi(u) du} = \frac{g(T(x), \theta)\pi(\theta)}{\int g(T(x), u)\pi(u) du} = \frac{g(t, \theta)\pi(\theta)}{\int g(t, u)\pi(u) du}. \tag{5.18}$$

Example 5.10. Consider an urn with Np red and $N(1-p)$ black balls. p is unknown but N is a known large number. Balls are drawn at random one by one with replacement, and the selection is stopped after n draws. For $i = 1, 2, \ldots, n$, let $Y_i = 1$ if the i-th ball drawn is red, and 0 otherwise. Then the Y_i are i.i.d. Ber(p). The likelihood function is $p^{\sum y_i}(1-p)^{n - \sum y_i}$. Now $X = \sum Y_i$ is sufficient for p. Suppose the prior distribution of p is Beta(α, β), with density

$$\pi(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha - 1}(1-p)^{\beta - 1}, \quad 0 < p < 1. \tag{5.19}$$

Here, $h(x \mid \theta) = f(x \mid \theta)\pi(\theta)$

$$= \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{x + \alpha - 1}(1-p)^{n - x + \beta - 1}. \tag{5.20}$$

Integration gives

$$\int_0^1 h(x \mid p) dp = \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{x + \alpha - 1}(1-p)^{n - x + \beta - 1} dp = \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + x)\Gamma(\beta + n - x)}{\Gamma(\alpha + \beta + n)}. \tag{5.21}$$

Thus the posterior density is

$$\pi(p \mid x) = \frac{h(x \mid p)}{\int_0^1 h(x \mid p)dp} = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, \quad 0 < p < 1. \tag{5.22}$$

Thus, $p \mid X = x \sim \text{Beta}(\alpha + x, \beta + n - x)$.

The HPD, *highest posterior density estimate* $\hat{p}_{\text{hpd}}$ is the value of p that maximizes the posterior density $\pi(p \mid x)$. In the example above, we have

$$\hat{p}_{\text{hpd}} = \frac{\alpha + x - 1}{\alpha + \beta + n - 2}. \tag{5.23}$$

The $\hat{p}_{\text{hpd}}$ is most probably the correct model, while the MLE is the parameter which most likely produced the data.

Definition 5.11. Let $\mathcal{F}$ denote a class of density functions $f(x \mid \theta)$. A class $\mathcal{P}$ of prior distributions is said to be conjugate for $\mathcal{F}$ if $\pi(\cdot \mid x) \in \mathcal{P}$ for all $f \in \mathcal{F}$ and $\pi \in \mathcal{P}$.

We discsuss *Jeffreys' prior*; let $f(x \mid \theta)$ be the model of $X \mid \theta$ for which $I(\theta)$ is the Fisher information. Then the Jeffreys' prior is defined as

$$\pi(\theta) = (I(\theta))^{1/2} = |I(\theta)|^{1/2}. \tag{5.24}$$

Definition 5.12. For $0 < \alpha < 1$, a $100(1-\alpha)\%$ *credible set* for θ is a subset $C \subseteq \Theta$ such that

$$P(C \mid X = x) = 1 - \alpha. \tag{5.25}$$

For the discrete case, the condition is relaxed to $P(C \mid X = x) \geq 1 - \alpha$.

5.2.1 Prediction of Future Observations

Suppose the data are $x_1, \ldots, x_n$ when $X_1, \ldots, X_n$ are i.i.d. $f(x \mid \theta)$, for example, $N(\mu, \sigma^2)$ with σ^2 known. We wish to predict the unobserved value of X_{n+1}.

Prediction by a single number $t(x_1, \ldots, x_n)$ amounts to considering prediction loss—

$$E[(X_{n+1} - t)^2 \mid \underline{x}] = E[(X_{n+1} - E[X_{n+1} \mid \underline{x}])^2 \mid \underline{x}] + (t - E[X_{n+1} \mid \underline{x}])^2 \tag{5.26}$$

which is minimum at $t = E[X_{n+1} \mid \underline{x}]$. To calculate the predictor we need to calculate the predictive distribution.

$$\pi(x_{n+1} \mid \underline{x}) = \int_\Theta \pi(x_{n+1} \mid \underline{x}, \theta)\pi(\theta \mid \underline{x})d\theta = \int f(x_{n+1} \mid \theta)\pi(\theta \mid \underline{x})d\theta. \tag{5.27}$$

Let $\mu(\theta) = \int_{-\infty}^{\infty} x f(x \mid \theta)dx$. It can be shown that

$$E[X_{n+1} \mid \underline{x}] = E[\mu(\theta) \mid \underline{x}] = \int_\Theta \mu(\theta)\pi(\theta \mid \underline{x})d\theta. \tag{5.28}$$

Index