INTRODUCTION TO STATISTICAL INFERENCE

Sangita Das, notes by Ramdas Singh

Third Semester

List of Symbols

Placeholder

Contents

<div align="center">

Chapter 1

SUFFICIENCY

</div>

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

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

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

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

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

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

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

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

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

$$f(x \mid \lambda) = \exp(x \log(\lambda) - \lambda - \log(x!))\mathbf{1}_A(x)$$
(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 set*s. 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 $\text{Exp}(\theta)$ 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

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

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 kth sample moment and the kth 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,*
>
> $$E[T(X)] = -\frac{d}{d\eta}d_0(\eta), \qquad and \qquad \mathrm{Var}(T(X)) = -\frac{d^2}{d\eta^2}d_0(\eta). \tag{2.13}$$

> **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 $E_\eta[T(X)]_{\eta=\hat\eta(x)} = T(x)$. ∎

Index