Chapter 2

Methods of Point Estimation

Contents

"All models are approximations. Assumptions, whether implied or clearly stated, are never exactly true. **All models are wrong, but some models are useful**. So the question you need to ask is not "Is the model true?" (it never is) but "Is the model good enough for this particular application?" — George Box, Alberto Luceno and Maria del Carmen Paniagua-Quinones.

Statistical inference is designed to make probabilistic statements about unknown quantities. These quantities summarize and reflect the information buried in data that people hope to know. For example the average lifetime of some electronic products, etc.. These quantities are also used to build *statistical models* that are used to predict how nature would behave in future. For example, people use historical data to do weather forecasting, and to predict stock prices.

A typical statistical problem consists of a random variable X of interest, its probability density function (pdf) $f(x)$, or probability mass function (pmf) $p(x)$ that is used to specify the distribution of X, and a set of sample $\{X_i : i = 1, 2, \ldots, m\}$ drawn independently from this distribution. Usually the application background restricts the

distribution of X to a specific family with only one or several undetermined parameters. For example, if we model the number of calls received by a telephone service, we probably will use a Poisson distribution with the parameter λ (which is also the mean of the distribution) undetermined. We use the methodology of statistical inference to guess/infer/find the unknwn parameter.

2.1 Sampling and statistics

Definition 2.1.1 (random sample)**.** *If the random variables $X_1, X_2, \ldots, X_m$ are independent and identically distributed (iid), then these random variables constitute a* ***random sample*** *of size m from the common distribution.*

Example 2.1.2. *For example, if we randomly select 20 light bulbs produced by a factory, and we do experiment to record the lifetimes $\{X_1, \ldots, X_{20}\}$ of the bulbs selected, then we get a random sample $\{X_i\}_{i=1}^{20}$. In application, since this kind of measurement would destroy the bulbs, nobody will do the experiment for the population which is all the bulbs produced by the factory, doing statistical inference based on this sample is very important for quality control, etc..*

Definition 2.1.3 (statistic, point estimator)**.** *Let $X_1, \ldots, X_m$ be a random sample from some specific probability distribution. Let $T = T(X_1, \ldots, X_m)$ be a function of the sample. Then T is called a* ***statistic****, or a* ***point estimator****.*

We continue the Example 2.1.2. For the purpose of finding the average lifetime, one may use the statistic $T_1(X_1, \ldots, X_m) = \frac{X_1 + \cdots + X_m}{m}$ which is simply the average of all the observations. One may also use a constant function $T_2(X_1, \ldots, X_m) = 1200(\text{hours})$ (sounds bizarre, but what if you trust your knowledge more than the experiments!). We will soon learn the statistical inference methodology to compare the quality of different statistics.

Remark 2.1.4. *An* ***estimator*** *is a function of the sample, while an* ***estimate*** *is the realized value of an estimator (that is, a number!) that is obtained when a sample is actually taken. Notationally, when a sample is taken, an estimator is a function of*

the random variables $X_1, \ldots, X_m$, while an estimate is a function fo the realized values $x_1, \ldots, x_m$.

We continue with Example 2.1.2. If we use the sample average as the estimator, and after finishing the experiment with 20 observed lifetime, the estimate is the average (for example, 1050 hours) of these 20 numbers.

2.2 Method of moments

The method of moments is an easy and direct way to construct estimators. However, it should be noted that this method sometimes does not produce good estimators[1].

Let $X_1, \ldots, X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, \theta_2, \ldots, \theta_k)$. *Method of moments* estimators are found by matching the first k *sample moments* to the corresponding k *population moments*, and solving the resulting system of simultaneous equations. More precisely, define

$$\begin{cases} m_1 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^1, & \mu_1 = E(X^1), \\ m_2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2, & \mu_2 = E(X^2), \\ \qquad\qquad \vdots & \\ m_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k, & \mu_k = E(X^k). \end{cases}$$

Usually the population moments $\mu_1, \ldots, \mu_k$ are functions of the parameters $\theta_1, \ldots, \theta_k$. We write the functions as $\mu_1(\theta_1, \ldots, \theta_k), \ldots, \mu_k(\theta_1, \ldots, \theta_k)$. Then we match m_i to μ_i for $i = 1, \ldots, k$ to give the following equation system of the unknown parameters

[1]We will discuss the quality of an estimator later.

$\theta_1, \ldots, \theta_k$:

$$\begin{cases} m_1 = \mu_1(\theta_1, \ldots, \theta_k), \\ m_2 = \mu_2(\theta_1, \ldots, \theta_k), \\ \qquad \vdots \\ m_k = \mu_k(\theta_1, \ldots, \theta_k). \end{cases} \tag{2.1}$$

We solve the equation system (2.1) to obtain the method of moments estimator (MME) $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ of the parameters $(\theta_1, \ldots, \theta_k)$.

Remark 2.2.1. *In statistics literature, an estimator of some parameter θ is usually written by adding a "hat" as $\hat{\theta}$.*

Example 2.2.2. *Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution with rate parameter λ. Find the moment estimator of λ.*

Recall that the population mean is $\mu_1 = 1/\lambda$, and the sample mean is $m_1 = \overline{X}$. We match $m_1 = \mu_1$ to give $1/\lambda = \overline{X}$, which implies $\hat{\lambda} = 1/\overline{X}$. □

Example 2.2.3. *Let $X_1, \ldots, X_n$ be a random sample from the* Gamma(α, β) *distribution. Find the moment estimator of α and β.*

Solution. For the Gamma (α, β) distribution, the mean is $\mu_1 = \alpha\beta$ and the variance is $\mu_2 - \mu_1^2 = \alpha\beta^2$, so that the second moment is $\mu_2 = \alpha\beta^2 + (\alpha\beta)^2$. We match the sample moments and the population moments to give

$$\begin{aligned} \overline{X} &= \alpha\beta, \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= (\alpha + \alpha^2)\beta^2. \end{aligned}$$

We solve the equations to give

$$\begin{aligned} \alpha = \overline{X}/\beta \Rightarrow \{\overline{X}/\beta + (\overline{X}/\beta)^2\}\beta^2 = m_2 \Rightarrow \beta\overline{X} = m_2 - \overline{X}^2 \\ \Rightarrow \hat{\beta} = \frac{m_2 - \overline{X}^2}{\overline{X}}, \qquad \hat{\alpha} = \frac{\overline{X}^2}{m_2 - \overline{X}^2}. \end{aligned}$$

□

Example 2.2.4 (for the tutorial)**.** *With the sample* $\{X_1, \ldots, X_n\}$*, use the method of moments to estimate* $\theta \in (0, \infty)$ *in the pdf.*

$$f(x;\theta) = (\theta^2 + \theta)x^{\theta-1}(1-x), \quad 0 \leq x \leq 1.$$

Example 2.2.5 (for the tutorial)**.** *Let* $X_1, \ldots, X_n$ *be a random sample from a* $N(\mu, \sigma^2)$ *distribution. Find the moment estimators of* μ *and* σ^2*.*

Example 2.2.6 (for the tutorial)**.** *Suppose that* $X_1 = 8.3$*,* $X_2 = 4.9$*,* $X_3 = 2.6$*, and* $X_4 = 6.5$ *is a random sample of size 4 from the two-parameter uniform pdf*

$$f(x;\theta_1,\theta_2) = \frac{1}{2\theta_2}, \quad \theta_1 - \theta_2 \leq x \leq \theta_1 + \theta_2.$$

Use the method of moments to estimate θ_1 *and* θ_2*.*

2.3 Maximum likelihood estimation

The following example shows the motivation of the maximum likelihood estimation.

Example 2.3.1 (maximum likelihood estimation, motivation)**.** *Suppose that you had just two observations* $x_1 = 6.5$*,* $x_2 = 7.5$ *from a normal population whose variance is* 1 *and the mean is known to be either* 7*, or* 10*. What would you choose as an estimate of the population mean, and why?*

We define the "likelihood", a quantity that measures the degree to which something is likely to happen, by the joint density function (or the joint probability mass function, in case of discrete random variables) values at the observations.

$$L(\mu) = f(x_1|\mu) \times f(x_2|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1-\mu)^2}{2}\right) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_2-\mu)^2}{2}\right)$$

Therefore, $L(7)$ measures that if the real mean $\mu = 7$, to what degree would the observations $x_1 = 6.5$ and $x_2 = 7.5$ be likely to happen; and $L(10)$ gives the degree to which the same observation is likely to happen. A simple computation shows that $L(7) = 0.12395$ and $L(10) = 1.52966 \times 10^{-5}$. This gives some credence to our intuition that it makes more sense to choose $\hat{\mu} = 7$.

Definition 2.3.2 (likelihood function)**.** *Let* $X_1, X_2, ..., X_n$ *be a random sample from a distribution characterized by* $f(x;\theta)$*, where* θ *is an unknown parameter. Let* $x_1, \ldots, x_n$ *be the observed values. The likelihood function is defined by*

$$L(\theta) = L(\theta; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i;\theta).$$

For the sake of simplicity in applications, people also use

Definition 2.3.3 (log-likelihood function)**.**

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

We see that if one parameter θ maximizes $L(\theta)$, it also maximizes $l(\theta)$. The maximum likelihood estimator is defined below.

Definition 2.3.4 (maximum likelihood estimator, **MLE**). *The maximum likelihood estimator* $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ *is defined to be the parameter that maximizes the likelihood function* $L(\theta)$ *(or equivalently, maximizes the log-likelihood function* $l(\theta)$*).*

Remark 2.3.5. *Note that the MLE,* $\hat{\theta}$ *is a function of the sample* $\hat{\theta}(X_1, \ldots, X_n)$*, and therefore it is a random variable. According to our previous Definition 2.1.3, the MLE* $\hat{\theta}$ *is a statistic, and is also referred to as a point estimator.*

Remark 2.3.6. *If the unknown distribution* f *is specified by more than one parameters,* $f = f(x; \theta_1, \ldots, \theta_k)$*, then the likelihood function is defined by*

$$L(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \ldots, \theta_k),$$

and the log-likelihood function is defined to be $l(\theta_1, \ldots, \theta_k) = \log L(\theta_1, \ldots, \theta_k)$*. The MLE is the vector* $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ *that maximizes* L *or* l*.*

How to find the MLE

The usual way to finding the maximizer of a function $L(\theta)$ is the following two steps:

1. Find θ such that

$$\frac{d}{d\theta} L(\theta) = 0; \tag{2.2}$$

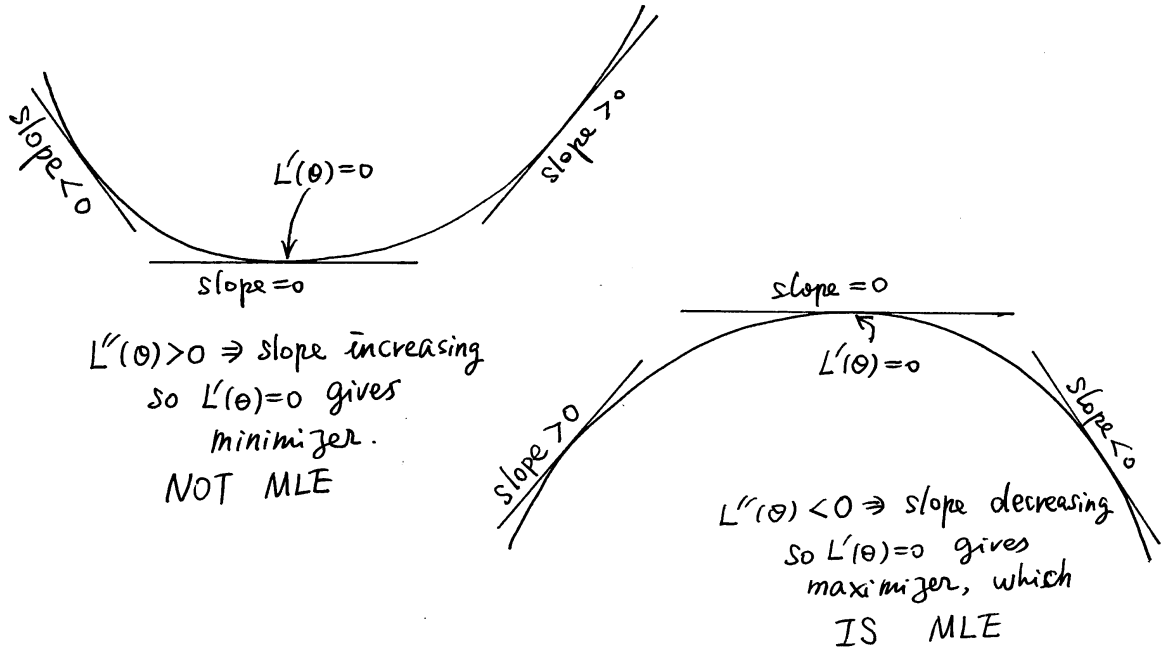
2. Verify this value of θ indeed is a maximizer by check that for this value of θ,

$$\frac{d^2}{d\theta^2} L(\theta) \leq 0. \tag{2.3}$$

Remark 2.3.7. *The solution to Equation (2.2) may not be unique.*

Remark 2.3.8. *Strictly speaking the above two steps will only find local maxima, while the MLE is the global maximum.*

Remark 2.3.9. *In both of the equations (2.2) and (2.3), the function* $L(\theta)$ *can be changed to* $l(\theta) = \log L(\theta)$ *since the logarithm function is strictly increasing.*



This figure is used to illustrate Equation(2.3).

Example 2.3.10. *Let $X_1, \ldots, X_n$ be independently sampled from $N(\theta, 1)$. Find the MLE of θ.*

Solution. The log-likelihood function (think, why we do not use likelihood function directly?) is

$$l(\theta) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\theta)^2}{2}\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\theta - x_i)^2.$$

One takes derivative to give

$$\frac{dl(\theta)}{d\theta} = -\sum_{i=1}^{n}(\theta - x_i).$$

Let $\frac{dl(\theta)}{d\theta} = 0$ to give $\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$. We verify that this is really the maximizer by looking at the second order derivative

$$\frac{d^2 l(\theta)}{d\theta^2} = -n < 0.$$

We replace the observation by the random sample to give the maximum likelihood

estimator (MLE),

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

□

Theorem 2.3.11 (Invariance property of MLEs)**.** *If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

Example 2.3.12 (we continue the Example 2.3.10)**.** *Let $X_1, \ldots, X_n$ be independently sampled from $N(\theta, 1)$. Find the MLE of θ^2.*

Solution. We have already found that $\hat{\theta} = \overline{X}$. So the MLE of θ^2 is $\hat{\theta}^2 = (\overline{X})^2$. □

Example 2.3.13 (for the tutorial)**.** *Let $X_1, \ldots, X_n$ denote a random sample from the distribution with pmf*

$$p(x;\theta) = \begin{cases} \theta^x(1-\theta)^{1-x}, & x = 0, 1 \\ 0, & elsewhere, \end{cases}$$

where $0 \leq \theta \leq 1$. Find the MLE of θ.

Example 2.3.14 (for the tutorial)**.** *Find the maximum likelihood estimator of the parameter μ in the* $\mathsf{Poisson}(\mu)$ *distribution. The Poisson pmf has the form*

$$f(x;\mu) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2, \ldots,$$

and the possible values of μ are $\mu \in (0, \infty)$.

Example 2.3.15 (for the tutorial)**.** *Suppose that a device consists of an original part and a backup that is automatically put into use when the original fails. Both parts fail at exponentially distributed times with a common rate parameter λ, and the failure times are independent of one another. Five copies of the device are observed, which fail at times 2.1, 3.2, 2.5, 4.6 and 3.8. Find the value of a maximum likelihood estimator for λ.*

Example 2.3.16 (for the tutorial)**.** *The problem of finding the MLE of θ in the uniform $(0, \theta)$ distribution requires a slightly different analysis. Let $X_1, X_2, ..., X_n$ be a random sample from this population, whose pdf is*

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

Example 2.3.17 (for the tutorial)**.** *The Weibull density with parameters λ and β is*

$$f(t;\lambda,\beta) = \beta\lambda^{\beta}t^{\beta-1}e^{-(\lambda t)^{\beta}}, \qquad t > 0$$

Let us try to find the MLE of β, assuming that it is known that $\lambda = 2$.

Example 2.3.18 (for the tutorial). *Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, $0 < \theta < \infty$, zero elsewhere. Find the MLE of θ.*

Example 2.3.19 (for the tutorial). *Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}$, $0 < x < \infty$, $0 < \theta < \infty$, zero elsewhere. Find the MLE of $P(X \leq 2)$.*

Example 2.3.20 (for the tutorial). *Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta) = e^{-(x-\theta)}$, $\theta \leq x < \infty$, $0 < \theta < \infty$, zero elsewhere. Find the MLE of θ.*

Example 2.3.21 (optional, not required)**.** *The maximum likelihood method applies as well to the estimation of a vector parameter. To illustrate, let us find the joint MLEs of* μ *and* σ^2 *in the normal distribution.*

Exercise 2.1

1. Let $X_1, \ldots, X_n$ be a random sample from the pdf $f(x;\theta) = \theta x^{-2}$, $0 < \theta \leq x < \infty$. Find the method of moments estimator (MME) of θ

2. Let $X_1, \ldots, X_n$ be a random sample from the pdf $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$, zero elsewhere. Find the method of moments estimator (MME) of θ.

3. Let $X_1, \ldots, X_n$ be a random sample from the pdf $f(x;\theta) = \theta(1+x)^{-(1+\theta)}$, $0 < x < \infty$, $\theta > 1$, zero elsewhere. Find the method of moments estimator (MME) of $1/\theta$.

4. Let $X_1, \ldots, X_n$ be a random sample from the pmf $f(x;\theta) = \frac{(\log\theta)^x}{\theta x!}$, $x = 0, 1, 2, \ldots$, $\theta > 1$, zero elsewhere. Find the method of moments estimator (MME) of θ.

5. (6.1.1) Let $X_1, \ldots, X_n$ be a random sample from a $\mathsf{Gamma}(\alpha = 3, \beta = \theta)$ distribution, $0 < \theta < \infty$. Determine the MLE of θ.

6. (6.1.4) Suppose $X_1, \ldots, X_n$ are iid with pdf $f(x;\theta) = 2x/\theta^2$, $0 < x \leq \theta$, zero elsewhere. Note this is a nonregular case. Find the MLE $\hat{\theta}$ for θ. Find a constant c so that $E(c\hat{\theta}) = \theta$.

7. (6.1.6) Let the table

x	0	1	2	3	4	5
Frequency	6	10	14	13	6	1

   represent a summary fo a sample of size 50 from a binomial distribution having $n = 5$. Find the MLE of $P(X \geq 3)$.

8. (6.1.8) Let the table

x	0	1	2	3	4	5
Frequency	7	14	12	13	6	3

   represent a summary of a random sample of size 55 from a Poisson distribution. Find the MLE of $P(X = 2)$.

9. (6.1.9)* Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p. If p is restricted so that we know that $\frac{1}{2} \leq p \leq 1$, find the MLE of this parameter.

10. (6.1.10)* Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ distribution, where σ^2 is fixed but $-\infty < \theta < \infty$.

    a. Show that the MLE of θ is $\overline{X}$.

    b. If θ is restricted by $0 \leq \theta < \infty$, show that the MLE of θ is $\hat{\theta} = \max\left\{0, \overline{X}\right\}$.

11. (6.1.12)* Let $X_1, \ldots, X_n$ be a random sample from a distribution with one of two pdfs. If $\theta = 1$ then $f(x;\theta = 1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $-\infty < x < \infty$. If $\theta = 2$, then $f(x;\theta = 2) = \frac{1}{\pi(1+x^2)}$, $-\infty < x < \infty$. Find the MLE of θ.

2.4 Supplementary notes on the moment generating functions (MGFs)

This section has no relation with the other part of the chapter. During the first lecture it was found that some students have not learned about the topic, therefore the topic is presented here.

Definition 2.4.1. *The **moment generating function** (MGF) of a random variable X is defined by*

$$M_X(t) = E[e^{tX}]$$

when the expectation exists.

Example 2.4.2. *Suppose $X \sim \mathsf{Poisson}(\lambda)$, find the MGF of X.*

Solution.

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t\lambda)^k}{k!} = e^{-\lambda} \exp(e^t\lambda).$$

The MGF exists for all $-\infty < t < \infty$. □

Theorem 2.4.3 (properties of MGF)**.** *Let rv's X and Y be independent. Let $M_X(t)$ and $M_Y(t)$ be their MGF respectively.*

1. $M_X^{(m)}(0) = E(X^m)$ *for every positive integer m.*
2. $M_{X+Y}(t) = M_X(t)M_Y(t)$
3. *If $M_X(t) \equiv M_Y(t)$, then X and Y have the same distribution.*
4. $M_{aX+b}(t) = e^{bt}M_X(at)$ *for any constants a and b.*

A sketch proof. One sees the point 1 by

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}E(e^{tX}) = E\left(\frac{d}{dt}e^{tX}\right) = E(Xe^{tX}),$$

which leads to

$$\frac{d}{dt}M_X(t)\Big|_{t=0} = E(Xe^{0\times X}) = E(X).$$

Similarly,

$$\begin{aligned} M_X^{(m)}(0) = \frac{d^m}{dt^m}M_X(t)\Big|_{t=0} &= \frac{d^m}{dt^m}E(e^{tX})\Big|_{t=0} = E\left(\frac{d^m}{dt^m}e^{tX}\right)\Big|_{t=0} \\ &= E\left(X^m e^{tX}\right)\big|_{t=0} = E(X^m). \end{aligned}$$

Point 2 is obtained by (recall that X and Y are independent, therefore E^{tX} and e^{tY} are independent.)

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX} \times e^{tY}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t).$$

Point 3 is quite involved and we skip it here. The proof of Point 4 is left as an exercise.

Example 2.4.4 (left as an exercise)**.** *Using MGF, prove that for any positive integer n and positive number θ,*

$$X \sim \mathsf{Gamma}(n, \theta) \Longleftrightarrow \frac{2}{\theta}X \sim \chi^2_{2n}.$$

2.5 Supplementary notes on asymptotic theories of MLE

Reference: Chapter 6, Theory of Point Estimation, by Erich L,. Lehmann.

Under some smoothness assumptions we shall in the present section prove the existence of asymptotically efficient estimators and provide a method for determining such estimators which, in many cases, leads to an explicit solution. We begin with the following assumptions:

(A0) The distributions P_θ of the observations are distinct (otherwise, θ cannot be estimated consistently)

(A1) The distributions P_θ have common support.

(A2) The observations are $X = (X_1, ..., X_n)$, where the X_i are iid with probability density $f(x_i|\theta)$.

(A3) The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.

Note: The true value of θ will be denoted by θ_0.

The joint density of the sample $L(\theta|\mathbf{x}) = \Pi_{i=1}^n f(x_i|\theta) = f(x_1|\theta)\cdots f(x_n|\theta)$, considered as a function of θ, plays a central role in statistical estimation, with a history dating back to the eighteenth century.

Theorem 2.5.1. *Under assumptions (A0)–(A2),*

$$P_{\theta_0}(L(\theta_0|\mathbf{X}) > L(\theta|\mathbf{X})) \to 1 \quad as \quad n \to \infty, \tag{2.4}$$

for any fixed $\theta \neq \theta_0$.

Proof. The inequality is equivalent to

$$\frac{1}{n}\sum \log[f(X_i|\theta)/f(X_1|\theta_0)] < 0.$$

By the law of large numbers, the left side tends in probability toward

$$E_{\theta_0} \log[f(X|\theta)/f(X|\theta_0)].$$

Since $-\log$ is strictly convex, Jensen's inequality shows that

$$E_{\theta_0} \log[f(X|\theta)/f(X|\theta_0)] < \log E_{\theta_0}[f(X|\theta)/f(X|\theta_0)] = 0,$$

and the result follows. □

By the above theorem, the density of X at the true θ_0 exceeds that at any other fixed θ with high probability when n is large. We do not know θ_0, but we can determine

the value $\hat{\theta}$ which maximizes the density of X, that is, which maximizes the likelihood function at the observed $\mathbf{X} = \mathbf{x}$. If this value exists and is unique, it is the maximum likelihood estimator (MLE) of θ. The MLE of $g(\theta)$ is defined to be $g(\hat{\theta})$. If g is 1:1 and $\xi = g(\theta)$, this agrees with the definition of $\hat{\xi}$ as the value of ξ. that maximizes the likelihood, and the definition is consistent also in the case that g is not 1:1.

The theorem also suggests that if the density of X varies smoothly with θ, the MLE of θ typically should be close to the true value of θ, and hence be a reasonable estimator.

Further reading: Theorem 3.7 and Theorem 3.10 of Lehmann's book.