Lecture 3  Point estimation

Tiejun Li

1 Basic methodology

From now on, we always denote the population (i.i.d. samples) as \( \{X_i\}_{i=1}^n \), the estimated parameters as \( \hat{\theta} \).

1.1 The method of moments

The method of moments is proposed by British statistician K. Pearson in 1894. It consists of equating the first few moments of a population to the corresponding moments of a sample, thus getting as many equations as are needed to solve for the unknown parameters.

Suppose the probability distribution of \( X \) is \( f(x; \theta) \), where \( \theta = (\theta_1, \ldots, \theta_m) \), and its first \( m \) moments \( \alpha_k \ (k = 1, \ldots, m) \) exists

\[
E X^k = \alpha_k(\theta).
\]

We estimate the parameters \( \theta \) by solving \( m \) equations

\[
\frac{1}{n} \sum_{i=1}^n X_i^k = \alpha_k(\hat{\theta}), \quad k = 1, \ldots, m
\]

Example 1. (Poisson distribution) Estimate the rate \( \lambda \) of \( \mathcal{P}(\lambda) \).

Solution. Because \( E X = \lambda \), we obtain the estimate of \( \lambda \) by the method of moments

\[
\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.
\]

Example 2. (Gaussian distribution) Estimate the mean and variance of Gaussian R.V. \( N(\mu, \sigma^2) \).

Solution. Because we have \( E X = \mu, E X^2 = \sigma^2 + \mu^2 \), we ask

\[
\bar{X} = \hat{\mu}, \quad \bar{X^2} = \hat{\sigma^2} + \hat{\mu^2}.
\]

Solving this system we obtain

\[
\hat{\mu} = \bar{X}, \quad \hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S_n^2 = (S_n^*)^2.
\]

From this derivation, we actually know that for any distribution, \( \bar{X} \) and \( (S_n^*)^2 \) are the moment estimate of mean and variance.

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China
Example 3. *(Uniform distribution)* Estimate the interval of uniform R.V. $U[\theta_1, \theta_2]$.

Solution. Because we have $E X = (\theta_1 + \theta_2)/2$, $\text{Var}(X) = (\theta_2 - \theta_1)^2/12$, we ask

$$\bar{X} = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}, \quad (S^*_n)^2 = \frac{(\hat{\theta}_2 - \hat{\theta}_1)^2}{12}.$$ 

We obtain

$$\hat{\theta}_1 = \bar{X} - \sqrt{3} S^*_n, \quad \hat{\theta}_2 = \bar{X} + \sqrt{3} S^*_n.$$

1.2 Maximum likelihood estimate (MLE)

The method of maximum likelihood is proposed by British statistician R.A. Fisher in 1912. He demonstrated the advantage of this method by showing that it yields sufficient estimators whenever they exist and that MLE are asymptotically minimum variance unbiased estimators.

Suppose the probability distribution of $X$ is $f(x; \theta)$, then we define the joint distribution of the population $\{X_i\}_{i=1}^n$ as

$$L(\theta|x) = \prod_{i=1}^n f(x_i|\theta),$$

which is called the likelihood function. The MLE is to obtain the parameters by solving the optimization problem

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta|x), \quad x = (X_1, \ldots, X_n).$$

Usually the MLE is obtained by maximizing the log-likelihood function $\ln L(\theta|x)$ since it has the same maximum with $L$. The Euler-Lagrange equation of this problem is

$$\frac{\partial \ln L(\theta|x)}{\partial \theta_i} = 0, \quad i = 1, \ldots, m$$

Solving this equation we obtain $\hat{\theta}$.

Example 4. *(Poisson distribution)* Estimate the rate $\lambda$ of $P(\lambda)$.

Solution. The log-likelihood function

$$\ln L(\lambda|x) = -n \lambda + \sum_{i=1}^n x_i \ln \lambda - \sum_{i=1}^n \ln(x_i!)$$

The MLE for Poisson distribution is still $\hat{\lambda} = \bar{X}$.

Example 5. *(Gaussian distribution)* Estimate the mean and variance of Gaussian R.V. $N(\mu, \sigma^2)$.

Solution. The log-likelihood function

$$\ln L(\lambda|x) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$ 

The MLE for Gaussian distribution is $\hat{\mu} = \bar{X}, \hat{\sigma}^2 = (S^*_n)^2$. 

2
Lemma 1. The MLE has the invariance property that if $\hat{\theta}$ is a MLE of $\theta$ and the function given by $g(\theta)$ is continuous, then $g(\hat{\theta})$ is also a MLE of $g(\theta)$.

Theorem 1. Suppose $T = T(X_1, \ldots, X_n)$ is a sufficient statistics of $\theta$, then the MLE $\hat{\theta}$ must be written as the function of $T$.

1.3 Bayesian estimation

In Bayesian estimations, the parameters are looked upon as random variable having prior distributions, usually reflecting the strength of one’s belief about the possible values that they can assume. More concretely, suppose we have the pdf $f(x|\theta)$ for $X$ with parameter $\theta$. The likelihood function is $L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$. The prior distribution is assumed as $\pi(\theta)$. Then the posterior distribution of the parameters is defined as

$$
\pi(\theta|x) = \frac{1}{m(x)} \pi(\theta)L(\theta|x),
$$

where

$$
m(x) = \int_{\Theta} \pi(\theta)L(\theta|x)d\theta
$$

is the marginal distribution of $(X_1, \ldots, X_n)$ which is independent of $\theta$. The posterior distribution reflects the contribution of the population to improve the knowledge on parameters.

In Bayesian statistical decision theory, we need a decision space $D$ and decision functions $d \in D$. Here $d$ is a function of $x$ in the product state space. Define the risk function

$$
R(\theta, d) = \mathbb{E}_\theta L(\theta, d(x)) = \int_{X} \mathcal{L}(\theta, d(x))L(\theta|x)dx,
$$

where $\mathcal{L}(\theta, d(x))$ is called the loss function. It is still a random function of $\theta$. Now we define the Bayesian risk

$$
R(d) = \mathbb{E}R(\theta, d) = \int_{\Theta} R(\theta, d(\theta))\pi(\theta)d\theta.
$$

Bayesian statistics is to choose $d$ such that the Bayesian risk is minimized

$$
d^* = \arg \inf_{d \in D} R(d)
$$

and estimated parameter $\hat{\theta} = d^*$.

Lemma 2. If the loss function is defined as

$$
\mathcal{L}(\theta, d(x)) = (\theta - d(x))^2,
$$

then the Bayesian estimate of $\theta$ is

$$
\hat{\theta} = \mathbb{E}(\theta|x) = \int_{\theta \in \Theta} \theta \pi(\theta|x)d\theta,
$$

i.e., the mean of the posterior distribution.
Proof. Under mild conditions we have
\[ R(d) = \int_\Theta \int_X L(\theta, d(x)) \pi(\theta) L(\theta|x) dx d\theta \]
\[ = \int_X \int_\Theta L(\theta, d(x)) \pi(\theta|x) m(x) d\theta dx \]
\[ = \int_X \int_\Theta (\theta - d(x))^2 \pi(\theta|x) m(x) d\theta dx \]

If for any \( x \), \( \int_\Theta (\theta - d(x))^2 \pi(\theta|x) d\theta \) achieves the minimum uniformly, \( R(d) \) will achieve minimum also. It is not difficult to prove that the optimal decision is
\[ d^*(x) = \mathbb{E}(\theta|x) = \int_\Theta \theta \pi(\theta|x) d\theta. \]

Lemma 3. If the loss function is defined as
\[ L(\theta, d(x)) = |\theta - d(x)|, \]
then the Bayesian estimate of \( \theta \) is the median of the posterior distribution.

Lemma 4. If \( \bar{X} \) is the mean of a random sample of size \( n \) from \( N(\mu, \sigma^2) \), where \( \sigma \) is known and the prior distribution \( \mu \sim N(\mu_0, \sigma_0^2) \). Then the posterior distribution of \( \mu \) given \( \bar{X} = \bar{x} \) is \( N(\mu_1, \sigma_1^2) \), where
\[ \mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}, \quad \frac{1}{\sigma_1^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}. \]

Some other choices of the parameter estimation in Bayesian statistics are maximizing the posterior distribution or direct sampling.

2 Theory

Definition 1. A statistic \( \hat{\theta} \) is an unbiased estimator of the parameter \( \theta \) if and only if \( \mathbb{E}\hat{\theta} = \theta \).

Clearly for Gaussian \( N(\mu, \sigma^2) \), exponential \( \mathcal{E}(\lambda) \), we have that \( \bar{X} \) is an unbiased estimator of \( \mu \) or \( \lambda \). But
\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = (S_n^*)^2 \]
is not an unbiased estimator. We have
\[ S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \]
is an unbiased estimator of \( \sigma^2 \).

Define the bias of the estimator \( \hat{\theta} \) as \( b(\theta) = \mathbb{E}\hat{\theta} - \theta \). If
\[ \lim_{n \to \infty} b(\theta) = 0, \]
the estimator is asymptotically unbiased.
2.1 Consistency

**Definition 2.** A statistic \( \hat{\theta} \) is a **consistent estimator** of the parameter \( \theta \) if and only if for each \( \epsilon > 0 \)

\[
\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0.
\]

Note that the consistency is an asymptotic property and it means the estimator converges to the parameters in probability.

It is a simple consequence of the Chebyshev’s theorem that

**Lemma 5.** If \( \hat{\theta} \) is an unbiased estimator and \( \text{Var}(\hat{\theta}) \to 0 \) \((n \to \infty)\), then \( \hat{\theta} \) is a consistent estimator of \( \theta \).

Note that it is a sufficient but not necessary condition.

2.2 Efficiency

Usually we have more than one unbiased estimator for a given parameter, we may choose the one whose sampling distribution has the smallest variance. We call it a minimum variance unbiased estimator (or best unbiased estimator).

Denote the pdf as \( f(x; \theta) \), here we only consider the one parameter case in this note. Define

\[
S(x, \theta) = \frac{\partial \ln f(x; \theta)}{\partial \theta}, \quad I(\theta) = \mathbb{E} S^2(X, \theta).
\]

\( I(\theta) \) is called the Fisher information. The Fisher information is simply the variance of the derivative of the log-likelihood function with respect to \( \theta \). We have the following famous Cramér-Rao inequality.

**Theorem 2.** (Cramér-Rao inequality) Suppose \( \hat{\theta} \) is an unbiased estimator of \( \theta \), we have

\[
\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}.
\]

**Proof.** Since \( \hat{\theta} \) is an unbiased estimator, we have

\[
\mathbb{E} \left( \hat{\theta}(X) - \theta \right) = \int \left( \hat{\theta}(x) - \theta \right) \cdot f(x; \theta) \, dx = 0.
\]

Differentiate on both sides we obtain

\[
\frac{\partial}{\partial \theta} \int \left( \hat{\theta}(x) - \theta \right) \cdot f(x; \theta) \, dx = \int \left( \hat{\theta}(x) - \theta \right) \frac{\partial f}{\partial \theta} \, dx - \int f \, dx = 0.
\]

By using the normalization condition of \( f \) we have

\[
\int \left( \hat{\theta} - \theta \right) f \frac{\partial \ln f}{\partial \theta} \, dx = 1.
\]

Factoring the integrand gives

\[
\int \left( \hat{\theta} - \theta \right) \sqrt{f} \left( \sqrt{f} \frac{\partial \ln f}{\partial \theta} \right) \, dx = 1.
\]

The Cauchy-Schwarz inequality gives

\[
\left( \int (\hat{\theta} - \theta)^2 f \, dx \right) \cdot \left( \int f \frac{\partial \ln f}{\partial \theta} \right)^2 \, dx \geq 1.
\]
Thus finally we obtain
\[
\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}.
\]

It is easy to obtain
\[
I_{X_1,X_2}(\theta) = I_{X_1}(\theta) + I_{X_2}(\theta)
\]
when \(X_1, X_2\) are two independent experiments. So
\[
\text{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}.
\]
for \(n\) samples. If the identity holds, we call this \(\hat{\theta}\) a minimum variance unbiased estimator of \(\theta\).

If \(\hat{\theta}_1\) and \(\hat{\theta}_2\) are two unbiased estimators of \(\theta\), and the variance of \(\hat{\theta}_1\) is less than that of \(\hat{\theta}_2\), we say that \(\hat{\theta}_1\) is relatively more efficient than \(\hat{\theta}_2\). We use the ratio
\[
\frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)}
\]
as a measure of the efficiency of \(\hat{\theta}_2\) relative to \(\hat{\theta}_1\).

### 2.3 Sufficiency

The concept sufficiency is proposed by R.A. Fisher in 1922. An estimator \(\hat{\theta}\) is said to be sufficient if it utilizes all the information in a sample relevant to the estimation of \(\theta\). That is, if all the knowledge about \(\theta\) from the sample is extracted. Mathematically we have

**Definition 3.** A statistic \(\hat{\theta} = T(X_1, \ldots, X_n)\) is a sufficient estimator of the parameter \(\theta\) if for any \(t\), the conditional probability distribution of \(X_1, \ldots, X_n\), given \(T(X_1, \ldots, X_n) = t\), is independent of \(\theta\).

We can explain the sufficient statistics as follows. The information about parameters \(\theta\) is contained in the samples because the joint distribution depends on \(\theta\). With condition that the sufficient statistics \(T = t\), the conditional distribution does not depend on \(\theta\), which means that the rest part information in the samples does not contain the information about \(\theta\). In another word, the whole information about \(\theta\) is contained in \(T\). Thus to make inference on \(\theta\), we only need \(T\).

**Example 6.** For \(X_i \sim B(p, q)\), show that \(\hat{p} = \bar{X}\) is a sufficient estimator of \(p\).

**Solution.** Given \(\bar{X} = \hat{p}\), we have the joint distribution
\[
f(x, \bar{x} = \hat{p}) = \prod_{i=1}^{n} p^{x_i} q^{1-x_i}, \quad x_i = 0 \text{ or } 1
\]
\[
= p^{n\hat{p}} q^{n(1-\hat{p})}.
\]
For the probability when the condition \(\bar{X} = \hat{p}\) holds, we have from \(B(n, p)\) for \(\sum_i X_i\)
\[
m(\bar{x} = \hat{p}) = \binom{n}{\hat{p}} p^{\hat{p}} q^{n-\hat{p}}.
\]
Thus the conditional distribution
\[ f(x|\hat{p}) = \frac{f(x, \hat{p})}{m(\hat{p})} = \frac{1}{C_n^{\hat{p}}} \]
which does not depend on \( p \).

Because it is very tedious to check whether a statistic is a sufficient estimator based on definition, it is usually easier to base it instead on the following **factorization theorem**.

**Theorem 3.** (Fisher-Neyman factorization theorem) \( T(X_1, \ldots, X_n) \) is a sufficient statistic iff the joint probability distribution can be factored so that
\[ f(x; \theta, T(x)) = g(t; \theta) \cdot h(x) \]
where \( g(t; \theta) \) depends only on \( t \) and \( \theta \), and \( h(x) \) does not depend on \( \theta \).

By Fisher-Neyman factorization theorem, we have that the information provided by a sufficient statistic \( T(X) \) is the same as that of the sample \( X \) because
\[ \frac{\partial \ln f(x; \theta, T(x) = t)}{\partial \theta} = \frac{\partial \ln g(t; \theta)}{\partial \theta} \]

### 3 Prior distributions in Bayesian statistics

The choice of the prior distribution in Bayesian statistics is a crucial step and its choice reflects the subjectivity. Different choices are utilized in realistic computations.

#### 3.1 Conjugate priors

Conjugate priors are common at the beginning of Bayesian statistics because it is easier in calculation. The concept, as well as the term “conjugate prior”, were introduced by Howard Raiffa and Robert Schlaifer in their work on Bayesian decision theory.

A class of prior probability distributions \( \pi(\theta) \) is said to be conjugate to a class of likelihood functions \( L(x|\theta) \) if the resulting posterior distributions \( \pi(\theta|x) \) are in the same family as \( \pi(\theta) \). For example, the Gaussian family is conjugate to itself (or self-conjugate) for the mean \( \mu \): if the likelihood function is Gaussian with known variance \( \sigma^2 \), choosing a Gaussian prior for mean \( \mu \) will ensure that the posterior distribution is also Gaussian. Note that the conjugate prior is only meaningful for a typical parameter of a typical distribution.

**Example 7.** Determine the conjugate priors for the variance \( \sigma^2 \) in \( N(\mu, \sigma^2) \) with known mean.

**Solution.** The likelihood function
\[ L(\theta|x) = \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \]
\[ \propto \left( \frac{1}{\sigma^2} \right)^{\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \]
Suppose $X \sim \Gamma(k, \theta)$, the pdf is
\[
\frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta}, \quad x > 0.
\]
We have that the pdf for $Y = X^{-1}$ is
\[
\frac{1}{\theta^k \Gamma(k)} \left(\frac{1}{y}\right)^{k+1} \exp \left(-\frac{1}{\theta y}\right), \quad y > 0.
\]
It is called inverse Gamma distribution, which is usually denoted as $IG(k, \theta)$. If the variance $\sigma^2$ satisfies $IG(k, \theta)$, then the posterior distribution of $\sigma^2$ is
\[
\pi(\sigma^2 | x) \propto \left(\frac{1}{\sigma^2}\right)^{k+1+\frac{n}{2}} \exp \left[-\frac{1}{\sigma^2} \left(\frac{1}{\theta} + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right)\right].
\]
It is still the inverse Gamma distribution $IG\left(k + n/2, \frac{1}{\theta} + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right)^{-1}$. The form
\[
\left(\frac{1}{\sigma^2}\right)^{k+1} \exp \left(-\frac{1}{\theta \sigma^2}\right)
\]
is called the kernel for $\sigma^2$.

Suppose the likelihood function is $L(\theta | x)$, $T = T(X_1, \ldots, X_n)$ is a sufficient statistic, then by factorization theorem
\[
L(x; \theta, T(x) = t) = g_n(t; \theta) \cdot h(x)
\]

**Theorem 4.** Suppose the function $f(\theta) \geq 0$, and
\[
0 < \int g_n(t; \theta) f(\theta) d\theta < \infty,
\]
then the set of functions
\[
D_f = \left\{ \frac{g_n(t; \theta) f(\theta)}{\int g_n(t; \theta) f(\theta) d\theta} : n = 1, 2, \ldots \right\}
\]
is the conjugate priors for the parameter $\theta$.

An incomplete list of the conjugate priors are shown in table 1.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameter</th>
<th>Conjugate Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B(n, p)$</td>
<td>$p$</td>
<td>$\beta$-distribution</td>
</tr>
<tr>
<td>$P(\lambda)$</td>
<td>$\lambda$</td>
<td>$\Gamma$-distribution</td>
</tr>
<tr>
<td>$\mathcal{E}(\lambda)$</td>
<td>$\lambda$</td>
<td>$\Gamma$-distribution</td>
</tr>
<tr>
<td>$N(\mu, \sigma^2)$</td>
<td>$\mu$</td>
<td>$N(\mu_1, \sigma_1^2)$</td>
</tr>
<tr>
<td>$N(\mu, \sigma^2)$</td>
<td>$\sigma^2$</td>
<td>$IG$-distribution</td>
</tr>
</tbody>
</table>

### 3.2 Maximum entropy priors

Maximum entropy priors and Jeffreys priors are two typical choices of the so-called *non-informative priors* in Bayesian statistics.
If some characteristics of the prior distributions (moments, etc.) are known and can be written as $K$ prior expectations

$$E_{\pi}[g_k(\theta)] = w_k,$$

a way to select a prior $\pi(\theta)$ satisfying these constraints is the maximum entropy method.

The Shannon entropy is defined as

$$S(\pi) = -\int \pi(\theta) \ln \pi(\theta) d\theta.$$

The distribution maximizing the entropy is of the form

$$\pi(\theta) = \frac{1}{Z} \exp \left( \sum_{k=1}^{K} \lambda_k g_k(\theta) \right),$$

where $Z = \int \exp \left( \sum_{k=1}^{K} \lambda_k g_k(\theta) \right) d\theta$, and $\{\lambda_k\}$ are Lagrange multipliers such that the $K$ constraints are satisfied.

### 3.3 Jeffreys priors

The Jeffrey’s prior is introduced by H. Jeffreys in 1961 through considering the measure transformation and the invariance principle.

- Step 1. Compute the log-likelihood function

$$l(\theta|x) = \sum_{i=1}^{n} \ln f(x_i|\theta);$$

- Step 2. Compute the Fisher information matrix

$$I(\theta) = \mathbb{E} \left( -\frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right), i,j=1,\ldots,p;$$

- Step 3. The Jeffreys prior is defined as

$$\pi(\theta) \propto \left[ \det I(\theta) \right]^{-\frac{1}{2}}.$$

**Example 8.** The Jeffreys prior for the parameter $p$ in Binomial distribution $B(n, p)$.

**Solution.** The pdf $f(k|p) = C_n^k p^k (1-p)^{n-k}$. The Fisher information

$$I(p) = -\mathbb{E} \left( \frac{\partial^2 \ln f(k|p)}{\partial p^2} \right) = \mathbb{E} \left( \frac{k}{p^2} + \frac{n-k}{(1-p)^2} \right) = \frac{n}{p(1-p)}.$$

Thus the Jeffreys prior is

$$\pi(p) \propto \left[ p(1-p) \right]^{-\frac{1}{2}} = Be\left(\frac{1}{2}, \frac{1}{2}\right).$$

### 3.4 Hierarchical prior distribution

When the hyper-parameters in the priors are difficult to be determined, one can apply prior distribution to these hyper-parameters again, which is called hierarchical prior distributions.
4 Homeworaks

• Compute the MLE for exponential distribution and the uniform distribution.

• Show that $\bar{X}$ is a minimum variance unbiased estimator of the mean $\mu$ of a Gaussian distribution.

• Show that the sample variance $S_n^2$ of a Gaussian distribution is a consistent estimator of $\sigma^2$.

• Prove that the conjugate prior for the parameter $p$ of $B(n, p)$ is Beta distribution.

References