Chapter 3: Estimation of Parameters

Mathematical Statistics

UIC

April 8, 2024

Overview

Statistical Inference

Statistical inference, or "learning", is the process of using data to infer the distribution that generated the data.

Basic Problem

We observe $X_1, \ldots, X_n \sim \pi$. We want to infer (or estimate, or learn) π or some features of π such as its mean.

Definition 3.1.1

A statistical model is a set of distributions or a set of densities (or PMFs) $\mathcal{F}$.

1. A parametric model is a set $\mathcal{F}$ that can be parameterized by a finite number of parameters.
2. A nonparametric model is a set $\mathcal{F}$ that cannot be parameterized by a finite set of parameters.

Example 3.1.2

1. If assume the data come from a normal distribution, then the model is

$$\mathcal{F} = \left\{ \pi\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \mu, \sigma^2 \in \mathbb{R} \right\},$$

   which is a two-parameter model. In $\pi\left(x|\mu,\sigma^2\right)$, x is a possible value of the random variable, whereas μ and σ^2 are parameters.

2. A nonparametric model:

$$\mathcal{F}_{\text{all}} = \{ \text{ all PDFs } \}$$

We will focus on parametric models. In general, a parametric model takes the form

$$\mathcal{F} = \{\pi(x|\theta), \quad \theta \in \Theta\}$$

where θ is an unknown parameter and Θ is the parameter space.

Remark: θ can be a vector, for instance, $\theta = \left(\mu, \sigma^2\right)$

Statistical Inference

Given a parametric model, $\mathcal{F} = \{\pi(x|\theta), \quad \theta \in \Theta\}$, the problem of inference is then to estimate (to learn) the parameter θ from the data.

Almost all problems in statistical inference can be identified as being one of three types: **point estimates**, **confidence intervals**, and **hypothesis testing**.

Three types of statistical inferences:

- **Point Estimation** refers to providing a single "best guess."

  Suppose $X_1, \ldots, X_n \sim \pi(x|\theta)$, where $\pi(x|\theta) \in \mathcal{F}$. A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

  $$\hat{\theta}_n = f(X_1, \ldots, X_n)$$

  Remember: θ is fixed but unknown, $\hat{\theta}_n$ is random since it depends on $X_1, \ldots, X_n$. We say that $\hat{\theta}_n$ is **unbiased** if

  $$\mathbb{E}\left[\hat{\theta}_n\right] = \theta$$

Cont'd

- A $100(1-\alpha)\%$ **Confidence Interval** for a parameter θ is a random interval $I_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ such that

  $$\mathbb{P}(\theta \in I_n) = 1 - \alpha$$

  In words: (a, b) traps θ with probability $1 - \alpha$.
  $(1 - \alpha)$ is called coverage of the confidence interval. In practice, $\alpha = 0.05$ is often used.

- In **Hypothesis Testing**, we start with some default theory, called a null hypothesis, and then ask if the data provide sufficient evidence to reject the theory. Otherwise, we fail to reject the null hypothesis.

Example 3.1.3

$X_1, \ldots, X_n \sim \text{Bernoulli}(p) : n$ independent coin flips. To test if the coin is fair, we test the null hypothesis $H_0 : p = 1/2$ against the alternative hypothesis $H_1 : p \neq 1/2$
It seems reasonable to reject H_0 if

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{2} \right| \quad \text{is large}$$

Summary

- A parametric model is a set $\mathcal{F}$ that can be parameterized by a finite number of parameters.
  - General parametric model:

    $$\mathcal{F} = \{\pi(x|\theta), \quad \theta \in \Theta\}$$

- A nonparametric model is a set $\mathcal{F}$ that cannot be parameterized by a finite set of parameters.
- Almost all problems in statistical inference can be identified as being one of three types:
  - Point Estimates
  - Confidence Intervals
  - Hypothesis Testing

Overview

Method of Moments: Problem Formulation

Suppose that
$$X_1, \ldots, X_n \sim \pi(x|\theta)$$
where $\theta \in \Theta$, and we want to estimate θ based on the data $X_1, \ldots, X_n$. The first method for constructing parametric estimators that we will study is called the method of moments.

- The estimators produced by this method are not optimal, but that are often easy to compute.
- They are also useful as starting values for other methods that require iterative numerical routines.

Method of Moments

Recall that the k^{th} moment of a probability distribution $\pi(x|\theta)$ is
$$\mu_k(\theta) = \mathbb{E}_\theta\left[X^k\right]$$
where $\mathbb{E}_\theta$ denotes expectation with respect to $\pi(x|\theta)$, i.e.
$$\mathbb{E}_\theta[f(X)] = \int f(x)\pi(x|\theta)\,\mathrm{d}x$$

If $X_1, \ldots, X_n$ are i.i.d from $\pi(x|\theta)$, then the k^{th} sample moment is defined as
$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

We can view $\hat{\mu}_k$ as an estimate of μ_k. Suppose that the parameter θ has k components:
$$\theta = (\theta_1, \ldots, \theta_k)$$

Method of Moments

Definition 3.2.1 (Method of Moments Estimator)

The **method of moments estimator** $\hat{\theta}$ is defined to be the value of θ such that
$$\begin{cases} \mu_1(\theta) = \hat{\mu}_1 \\ \mu_2(\theta) = \hat{\mu}_2 \\ \quad \cdots\cdots \\ \mu_k(\theta) = \hat{\mu}_k \end{cases} \tag{1}$$

- System (1) is a system of k equations with k unknowns: $\theta_1, \ldots, \theta_k$
- The solutions of this system $\hat{\theta}$ is the method of moments estimate of the parameter θ.

Example 3.2.2 (Bernoulli)

Let $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$. Find the method of moments estimate of the parameter p.

Example 3.2.3 (Normal)

Let $X_1, \dots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Find the method of moments estimates of μ and σ^2.

Consistency of the MoM estimator

Question: How good is the estimator $\hat{\theta}$ obtained by the method of moments?

Definition 3.2.4 (Consistency)

Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be consistent if

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$$

That is, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \to 0 \quad \text{as} \quad n \to \infty$$

Theorem 3.2.5

The method of moments estimate is consistent.

Summary

- If $X_1, \dots, X_n \sim \pi(x|\theta)$, then the method of moments estimate $\hat{\theta}$ of $\theta = (\theta_1, \dots, \theta_k)$ is the solution of

$$\begin{cases} \mu_1(\theta) = \hat{\mu}_1 \\ \mu_2(\theta) = \hat{\mu}_2 \\ \quad\vdots \\ \mu_k(\theta) = \hat{\mu}_k \end{cases}$$

  where

  - $\mu_k(\theta)$ is the k^{th} moment
  $$\mu_k(\theta) = \mathbb{E}_\theta\left[X^k\right]$$
  - $\hat{\mu}_k$ is the k^{th} sample moment
  $$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- The method of moments estimate $\hat{\theta}$ is a consistent estimate of θ.

Overview

The most common method for estimating parameters in a parametric model is the method of maximum likelihood.

Suppose $X_1, \dots, X_n$ are i.i.d. from $\pi(x|\theta)$.

Definition 3.3.1 (Likelihood Function)

The **likelihood function** is defined by

$$\boxed{\mathcal{L}(\theta) = \prod_{i=1}^{n} \pi\left(X_i|\theta\right)}$$

Important Remarks:

- The likelihood function is just the joint pdf/pmf of the data, except that we treat it as a function of the parameter θ.
- Thus, $\mathcal{L} : \Theta \to [0, \infty)$
- The likelihood function is not a density function: it is not true that $\mathcal{L}$ integrates to one, i.e $\int_\Theta \mathcal{L}(\theta)\, \mathrm{d}\theta \neq 1$.

Definition 3.3.2 (The Maximum Likelihood Estimate)

The **maximum likelihood estimate (MLE)** of θ, denoted $\hat{\theta}_{\mathrm{MLE}}$, is the value of θ that maximizes the likelihood $\mathcal{L}(\theta)$

$$\hat{\theta}_{\textsf{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$$

$\hat{\theta}_{\textsf{MLE}}$ makes the observed data $X_1, \dots, X_n$ "most probable" or "most likely"

Important Remark:
Rather than maximizing the likelihood itself, it is often easier to maximize its natural logarithm (which is equivalent since the log is a monotonic function). The log-likelihood is

$$\boxed{l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \pi\left(X_i|\theta\right)}$$

Example 3.3.3 (Bernoulli)

$X_1, \dots, X_n \sim \mathrm{Bernoulli}(p)$. Find the MLE of p.

- Answer:

$$\hat{p}_{\textsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$$

- In this example, $\hat{p}_{\mathrm{MLE}} = \hat{p}_{\mathrm{MoM}}$

Example 3.3.4 (Normal)

$X_1, \dots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Find the MLEs of μ and σ^2.

- Answer:

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \qquad \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

- Again, in this example, MLEs are the same as the MoM estimates.

Properties of MLE

Under certain conditions on the model

$$\mathcal{F} = \{\pi(x|\theta), \quad \theta \in \Theta\}$$

(under some smoothness conditions of π), the MLE $\hat{\theta}_{\text{MLE}}$ possesses many attractive properties that make it an appealing choice of estimate.

Main properties of the MLE:

- MLE is consistent:
$$\hat{\theta}_{\text{MLE}} \xrightarrow{\mathbb{P}} \theta_0$$
where θ_0 denotes the true value of θ.
- MLE is equivariant: if $\hat{\theta}_{\text{MLE}}$ is the MLE of $\theta \Rightarrow f\left(\hat{\theta}_{\text{MLE}}\right)$ is the MLE of $f(\theta)$.
- MLE is asymptotically optimal: the MLE has the smallest variance for large sample sizes n.

Properties of MLE

Main properties of the MLE (cont'd):

- MLE is asymptotically normal:
$$\hat{\theta}_{\text{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$
where
$$I(\theta) \overset{\text{def}}{=} \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\pi(X|\theta)\right)^2\right] = \int\left(\frac{\partial}{\partial\theta}\log\pi(x|\theta)\right)^2\pi(x|\theta)\,\mathrm{d}x$$
    - $I(\theta)$ is called Fisher Information.
- MLE is asymptotically unbiased:
$$\lim_{n\to\infty}\mathbb{E}\left[\hat{\theta}_{\text{MLE}}\right] = \theta_0$$

Example: when MoM and MLE produce different estimates

Example 3.3.5 (Uniform)

Let $X_1, \ldots, X_n \sim U(0, \theta)$. Find the MoM estimate and MLE of θ.

- Answer:
$$\hat{\theta}_{\text{MoM}} = 2\overline{X}_n \quad \hat{\theta}_{\text{MLE}} = X_{(n)}$$
- In this example, the MLE and MoM estimate are different.

Summary

- The Likelihood Function:
$$\mathcal{L}(\theta) = \prod_{i=1}^{n}\pi(X_i|\theta) \quad X_1, \ldots, X_n \sim \pi(x|\theta)$$
- The Maximum Likelihood Estimate:
$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta}\mathcal{L}(\theta) = \arg\max_{\theta\in\Theta}\log\mathcal{L}(\theta)$$
- MLE is consistent, equivariant, asymptotically optimal, asymptotically normal, and asymptotically unbiased.
- Examples: Bernoulli (p), $N(\mu, \sigma^2)$, and $U(0, \theta)$.

Overview

Confidence Interval

Recall the definition of a confidence interval (see also Definition 2.4.12 and Theorem 2.4.13):

Definition 3.4.1 (Confidence Interval)

A $100(1-\alpha)\%$ **confidence interval** for a parameter θ is a *random* interval calculated from the sample,

$$X_1, \dots, X_n \sim \pi(x|\theta)$$

which contains θ with probability $1-\alpha$.

There are three methods for constructing confidence intervals using MLEs $\hat{\theta}_{\text{MLE}}$:

- Exact Method
- Approximate Method
- Bootstrap Method

Exact Method. Example: Normal distribution $\mathcal{N}(\mu, \sigma^2)$

Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then the MLEs for μ and σ^2 are (Example 3.3.4):

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \qquad \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

- A confidence interval for μ is based on the following fact (Theorem 1.7.29):

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \sim t_{n-1}$$

where S_n^2 is the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}\hat{\sigma}^2_{\text{MLE}}$

Result

A $100(1-\alpha)\%$ confidence interval for μ is

$$\hat{\mu}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n-1}}\hat{\sigma}_{\mathrm{MLE}} t_{n-1}(\alpha/2)$$

where $t_{n-1}(\alpha)$ is the point beyond which the t-distribution with $(n-1)$ degrees of freedom has probability α.

Exact Method. Example: Normal distribution $\mathcal{N}(\mu, \sigma^2)$

- A confidence interval for σ^2 is based on the following fact (Theorem 1.7.29):

$$\frac{(n-1)\, S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Result

A $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right)$$

where $\chi^2_{n-1}(\alpha)$ is the point beyond which the χ^2-distribution with $(n-1)$ degrees of freedom has probability α.

Remark:
The main drawback of the exact method is that in practice the sampling distributions like t_{n-1} and χ^2_{n-1} in our example are unknown.

Approximate Method

One of the most important properties of MLE is that it is asymptotically normal:

$$\hat{\theta}_{\text{MLE}} \rightarrow \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right), \quad \text{as } n \rightarrow \infty$$

where $l(\theta_0)$ is Fisher information

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\pi(X|\theta)\right)^2\right]$$

Since the true value θ_0 is unknown, we will use $I\left(\hat{\theta}_{\text{MLE}}\right)$ instead of $I(\theta_0)$:

Result

An **approximate** $100(1-\alpha)\%$ confidence interval for θ_0 is

$$\hat{\theta}_{\text{MLE}} \pm \frac{z_{\alpha/2}}{\sqrt{nI\left(\hat{\theta}_{\text{MLE}}\right)}}$$

where z_α is the point beyond which the standard normal distribution has probability α.

Approximate Method. Example: Bernoulli (p)

Example 3.4.2 (Bernoulli (p))

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Find an approximate confidence interval for p

- Answer:

$$\overline{X}_n \pm z_{\alpha/2}\sqrt{\frac{\overline{X}_n\left(1-\overline{X}_n\right)}{n}}$$

Bootstrap Method

Suppose $\hat{\theta}$ is an estimate of a parameter θ, the true unknown value of which is θ_0. $\hat{\theta}$ can be any estimate, not necessarily MLE,

$$X_1, \ldots, X_n \sim \pi(x|\theta) \quad \hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

Define a new random variable

$$\Delta = \hat{\theta} - \theta_0$$

- Step 1: Assume (for the moment) that the distribution of Δ is known. Let (as before) q_α be the number such that $\mathbb{P}(\Delta > q_\alpha) = \alpha$. Then

$$\mathbb{P}\left(q_{1-\frac{\alpha}{2}} \leq \hat{\theta} - \theta_0 \leq q_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

And therefore a $100(1-\alpha)\%$ confidence interval for θ_0 is

$$\left(\hat{\theta} - q_{\frac{\alpha}{2}}, \hat{\theta} - q_{1-\frac{\alpha}{2}}\right)$$

The problem is that the distribution of Δ is unknown and, therefore, q_α are unknown.

Bootstrap Method

- Step 2: Assume that the distribution of Δ is not known, but θ_0 is known. Then we can approximate the distribution of Δ as follows:

$$\begin{gathered} X_1^{(1)}, \ldots, X_n^{(1)} \sim \pi(x|\theta_0) \rightsquigarrow \hat{\theta}^{(1)} - \theta_0 = \Delta^{(1)} \\ X_1^{(2)}, \ldots, X_n^{(2)} \sim \pi(x|\theta_0) \rightsquigarrow \hat{\theta}^{(2)} - \theta_0 = \Delta^{(2)} \\ \vdots \\ X_1^{(B)}, \ldots, X_n^{(B)} \sim \pi(x|\theta_0) \rightsquigarrow \hat{\theta}^{(B)} - \theta_0 = \Delta^{(B)} \end{gathered}$$

From these realizations $\Delta^{(1)}, \ldots, \Delta^{(B)}$ of Δ we can approximate the distribution of Δ by its empirical distribution, and, therefore, we can approximate q_α. The problem is that θ_0 is not known!

Bootstrap Method

- Step 3: **Bootstrap strategy**: Use $\hat{\theta}$ instead of θ_0.

$$
\begin{aligned}
X_1^{(1)},\dots,X_n^{(1)} &\sim \pi\left(x|\hat{\theta}\right) \rightsquigarrow \hat{\theta}^{(1)} - \hat{\theta} \approx \Delta^{(1)} \\
X_1^{(2)},\dots,X_n^{(2)} &\sim \pi\left(x|\hat{\theta}\right) \rightsquigarrow \hat{\theta}^{(2)} - \hat{\theta} \approx \Delta^{(2)} \\
&\vdots \\
X_1^{(B)},\dots,X_n^{(B)} &\sim \pi\left(x|\hat{\theta}\right) \rightsquigarrow \hat{\theta}^{(B)} - \hat{\theta} \approx \Delta^{(B)}
\end{aligned}
$$

Distribution of Δ is approximated from realizations $\Delta^{(1)},\dots,\Delta^{(B)}$.

Remark: $\hat{\theta}^{(i)}$ is the estimate of θ that is obtained from $X_1^{(i)},\dots,X_n^{(i)}$ by the same method (for example, MLE) as $\hat{\theta}$ was obtained from $X_1,\dots,X_n$.

Summary

- Three methods for constructing confidence intervals using MLEs:
- Exact method provides exact confidence intervals, but it's hard to use in practice
  - Example: $X_1,\dots,X_n \sim \mathcal{N}\left(\mu,\sigma^2\right)$

$$
\mu : \quad \hat{\mu}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n-1}}\hat{\sigma}^2_{\mathrm{MLE}} t_{n-1}(\alpha/2)
$$

$$
\sigma^2 : \quad \left(\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right)
$$

- Approximate method provides an approximate confidence interval for θ_0, which is constructed using asymptotic properties of MLE:

$$
\hat{\theta}_{\text{MLE}} \pm \frac{z_{\alpha/2}}{\sqrt{nI\left(\hat{\theta}_{\text{MLE}}\right)}}
$$

- Bootstrap method provides an approximate confidence interval. Bootstrap is the most popular method in practice since it is easy to implement.

Overview

Example: Gaussian Model

Suppose that:

- $X_1,\dots,X_n \sim \mathcal{N}\left(\mu,\sigma^2\right)$, true values: $\mu = 1$ and $\sigma = 2$
- Exact Confidence Intervals:

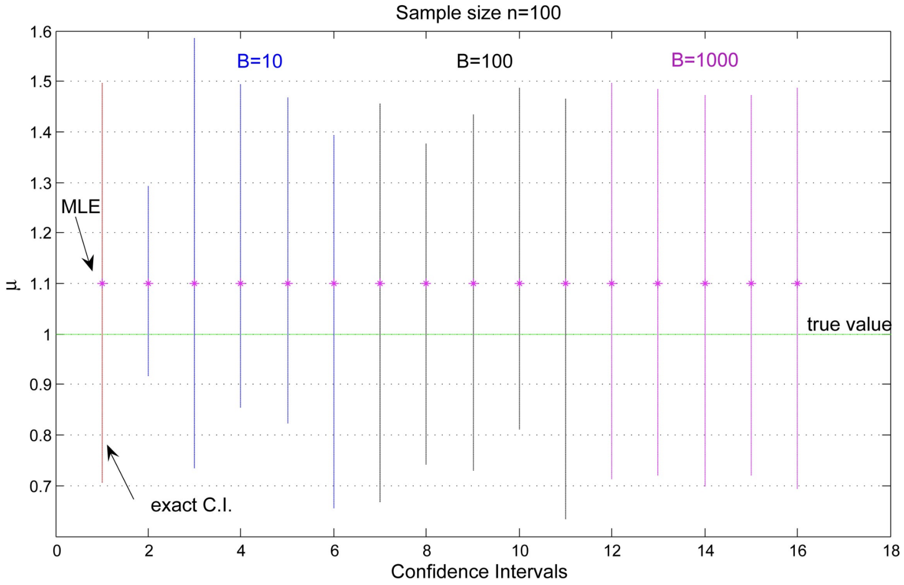
$$
\mu : \quad \hat{\mu}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n-1}}\hat{\sigma}_{\mathrm{MLE}} t_{n-1}(\alpha/2) \qquad \sigma^2 : \quad \left(\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right)
$$

```
%---- Data:
mu0=1;                     % true mean
sigma0=2;                  % true sigma
n=100;                     % sample size;
X=mu0+sigma0*randn(1,n); % data
%---- MLEs:
mu_mle=mean(X);
sigma_mle=std(X,1);
%---- Level of Confidence:
alpha=0.05;                % 100(1-alpha) CI
%---- Exact Confidence Intervals:
CImu_exact=[mu_mle-sigma_mle*tinv(1-alpha/2,n-1)/sqrt(n-1),
mu_mle+sigma_mle*tinv(1-alpha/2,n-1)/sqrt(n-1)];
CIsigma_exact=[sqrt(n*sigma_mle^2/chi2inv(1-alpha/2,n-1)),
sqrt(n*sigma_mle^2/chi2inv(alpha/2,n-1))];
%[phat,pci] = mle(X);
```

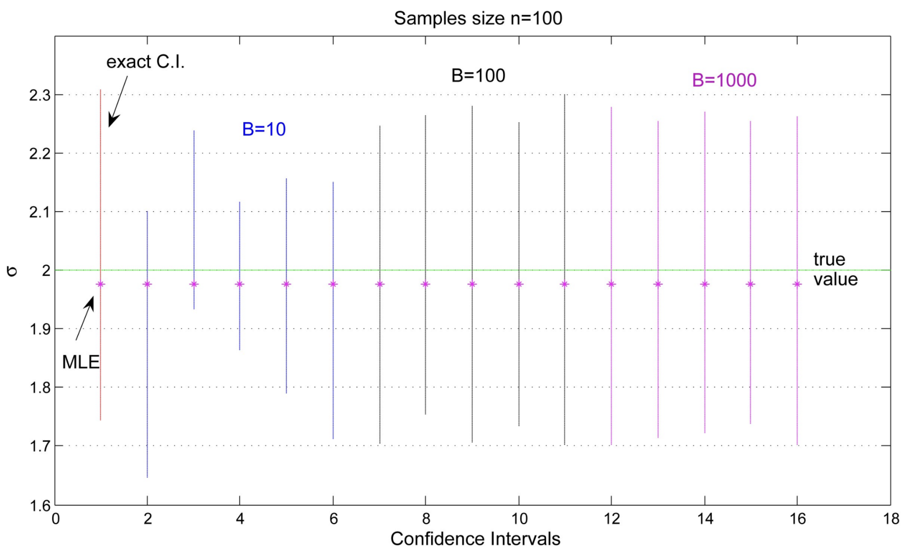
Bootstrap

```
%---- Bootstrap Confidence Intervals:
B=10;     % number of the bootstrap samples
for i=1:B
    Z(i,:)=mu_mle+sigma_mle*randn(1,n); % "bootstrap data"
    mu_b(i)=mean(Z(i,:));               % MLE from b-data
    sigma_b(i)=std(Z(i,:),1);           % MLE from b-data
    Delta_mu(i)=mu_b(i)-mu_mle;
    Delta_sigma(i)=sigma_b(i)-sigma_mle;
end
CImu_bootstrap=[mu_mle-quantile(Delta_mu,1-alpha/2),
    mu_mle-quantile(Delta_mu,alpha/2)];
CIsigma_bootstrap=[sigma_mle-quantile(Delta_sigma,1-alpha/2),
    sigma_mle-quantile(Delta_sigma,alpha/2)];
```

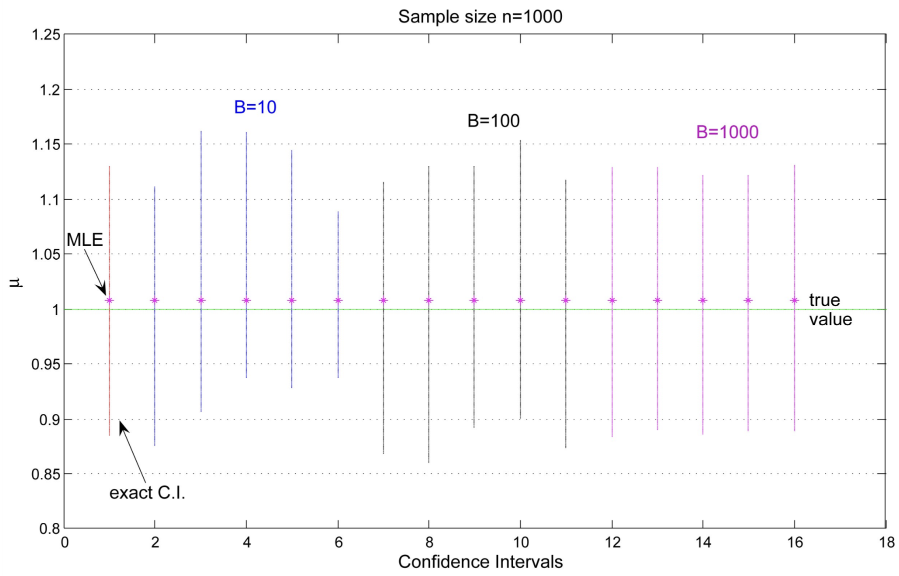

Confidence Intervals for μ when $n = 100$



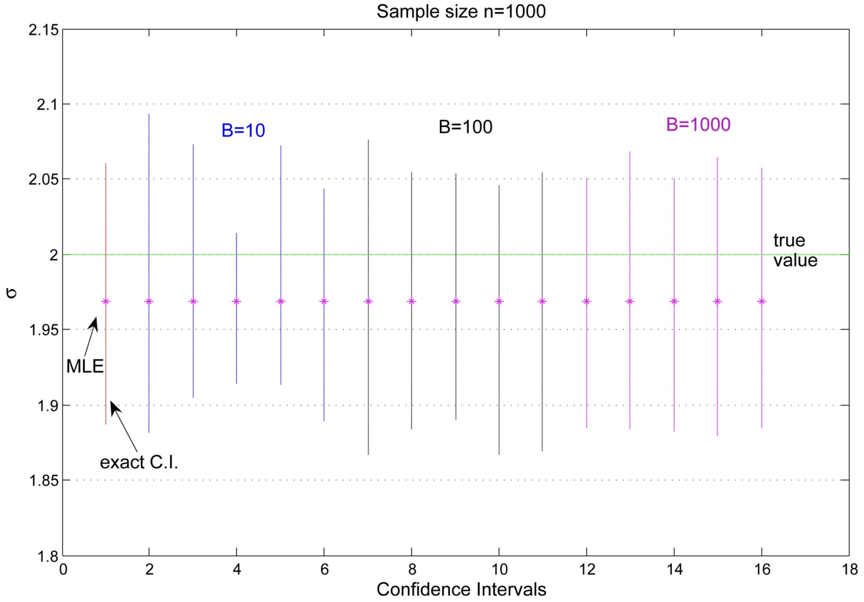
Confidence Intervals for σ when $n = 100$



Confidence Intervals for μ when $n = 1000$



Confidence Intervals for σ when $n = 1000$



Overview

Measure of Efficiency: Mean Squared Error

In most estimation problems, there are many possible estimates $\hat{\theta}$ of θ. For example, the MoM estimate $\hat{\theta}_{\text{MoM}}$ or the MLE estimate $\hat{\theta}_{\text{MLE}}$.

Question: How would we choose which estimate to use?

Qualitatively, it is reasonable to choose that estimate whose distribution is most highly concentrated about the true parameter value θ_0. To make this idea work, we need to define a quantitative measure of such concentration.

Definition 3.5.1 (Mean-squared Error)

The **mean squared error** of $\hat{\theta}$ as an estimate of θ_0 is

$$\boxed{\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)^2\right]}$$

- The mean squared error can be also written as follows:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}] + \underbrace{\left(\mathbb{E}(\hat{\theta}) - \theta_0\right)^2}_{\text{squared bias}}$$

- If $\hat{\theta}$ is unbiased, then $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}]$.

Cramer-Rao Inequality

- Given two unbiased estimates, $\hat{\theta}$ and $\tilde{\theta}$, the **efficiency** of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be

$$\mathrm{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\mathrm{Var}(\tilde{\theta})}{\mathrm{Var}(\hat{\theta})}$$

- $\hat{\theta}$ is more efficient than $\tilde{\theta}$ $\Leftrightarrow$ $\mathrm{eff}(\hat{\theta}, \tilde{\theta}) > 1$
- In general, the mean squared error is a measure of efficiency of an estimate:

the smaller $\mathrm{MSE}(\hat{\theta})$, the better the estimate $\hat{\theta}$

Theorem 3.5.2 (Cramer-Rao Inequality)

Let $X_1, \ldots, X_n$ be i.i.d. from $\pi(x|\theta)$. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be any unbiased estimate of a parameter θ whose true values is θ_0. Then, under smoothness assumptions on $\pi(x|\theta)$,

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}$$

Cramer-Rao: $\boxed{\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}}$

Important Remarks:

- $\hat{\theta}$ can't have arbitrary small MSE
- The Cramer-Rao inequality gives a lower bound on the variance of any unbiased estimate.

Definition 3.5.3 (Efficient)

An unbiased estimate whose variance achieves this lower bound is said to be **efficient**.

Recall that MLE is asymptotically Normal: $\hat{\theta}_{\text{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$

- Therefore, MLE is asymptotically efficient
- However, for a finite sample size n, MLE may not be efficient
- MLEs are not the only asymptotically efficient estimates.

Example: Poisson Distribution

Recall: The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events k occurring in a fixed interval of time if these events occur with a known average rate λ and independently of the time since the last event.

$$\mathbb{P}(X = k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \mathbb{E}[X] = \lambda \quad \mathrm{Var}[X] = \lambda$$

Example 3.5.4 (Poisson)

Let $X_1, \ldots, X_n \sim \mathrm{Pois}(\lambda)$.

- Find the MLE of λ
- Show that $\hat{\lambda}_{\text{MLE}}$ is efficient.

- The theorem does not exclude the possibility that there is a biased estimator of λ that has a smaller MSE than $\hat{\lambda}_{\text{MLE}}$

Summary

- Mean squared error is a measure of efficiency of an estimate

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)^2\right]$$

- If $\hat{\theta}$ is unbiased, then

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}]$$

- Cramer-Rao Inequality:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}$$

- An unbiased estimate whose variance achieves this lower bound is said to be efficient
- Any MLE is asymptotically efficient (as $n \to \infty$)
- Example: if $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$, then $\hat{\lambda}_{\mathrm{MLE}}$ is efficient