

Multivariate Statistics

Lecture 06

Fudan University

1 Efficiency

Outline

- 1 Efficiency
- 2 Consistency

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality
- 4 Decision Theory

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality
- 4 Decision Theory
- 5 The Biased Estimator

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality
- 4 Decision Theory
- 5 The Biased Estimator
- 6 Chi-Squared Distribution

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality
- 4 Decision Theory
- 5 The Biased Estimator
- 6 Chi-Squared Distribution

Efficiency

If a p -component random vector \mathbf{y} has mean vector $\mathbb{E}[\mathbf{y}] = \boldsymbol{\nu}$ and covariance matrix $\mathbb{E}[(\mathbf{y} - \boldsymbol{\nu})(\mathbf{y} - \boldsymbol{\nu})^\top] = \boldsymbol{\Psi} \succ \mathbf{0}$, then

$$\left\{ \mathbf{z} : (\mathbf{z} - \boldsymbol{\nu})^\top \boldsymbol{\Psi}^{-1} (\mathbf{z} - \boldsymbol{\nu}) = p + 2 \right\}$$

is called the concentration ellipsoid of \mathbf{y} .

Let $\boldsymbol{\theta}$ be a vector of p parameters in a distribution, and let \mathbf{t} be a vector of unbiased estimators (that is, $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$) based on N observations from that distribution with covariance matrix $\boldsymbol{\Psi}$. Then the ellipsoid

$$\left\{ \mathbf{z} : N(\mathbf{z} - \boldsymbol{\theta})^\top \mathbb{E} \left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right] (\mathbf{z} - \boldsymbol{\theta}) = p + 2 \right\}$$

lies entirely within the ellipsoid of concentration of \mathbf{t} , where f is the density of the distribution (or probability function) with respect to the components of $\boldsymbol{\theta}$.

The ellipsoid

$$\left\{ \mathbf{z} : N(\mathbf{z} - \boldsymbol{\theta})^\top \mathbb{E} \left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right] (\mathbf{z} - \boldsymbol{\theta}) = p + 2 \right\}$$

lies entirely within the ellipsoid of concentration of \mathbf{t}

$$\left\{ \mathbf{z} : (\mathbf{z} - \boldsymbol{\theta})^\top \left(\mathbb{E} [(\mathbf{t} - \boldsymbol{\theta})(\mathbf{t} - \boldsymbol{\theta})^\top] \right)^{-1} (\mathbf{z} - \boldsymbol{\theta}) = p + 2 \right\},$$

that is

$$\left(N \mathbb{E} \left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right] \right)^{-1} \preceq \mathbb{E} [(\mathbf{t} - \boldsymbol{\theta})(\mathbf{t} - \boldsymbol{\theta})^\top].$$

Efficiency

Let $\boldsymbol{\theta}$ be a vector of p parameters in a distribution, and let \mathbf{t} be a vector of unbiased estimators (that is, $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$) based on N observations from that distribution with covariance matrix $\boldsymbol{\Psi}$. Then the ellipsoid

$$\left\{ \mathbf{z} : N(\mathbf{z} - \boldsymbol{\theta})^\top \mathbb{E} \left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right] (\mathbf{z} - \boldsymbol{\theta}) = p + 2 \right\} \quad (1)$$

lies entirely within the ellipsoid of concentration of \mathbf{t} , where f is the density of the distribution (or probability function) with respect to the components of $\boldsymbol{\theta}$.

- 1 If the ellipsoid (1) is the ellipsoid of concentration of \mathbf{t} , then \mathbf{t} is said to be efficient.
- 2 In general, the ratio of the volume of (1) to that of the ellipsoid of concentration defines the efficiency of \mathbf{t} .

Consider the case of the multivariate normal distribution.

- ① If $\theta = \mu$, then $\bar{\mathbf{x}}$ is efficient.
- ② If θ includes both μ and Σ , then $\bar{\mathbf{x}}$ and \mathbf{S} have efficiency $((N-1)/N)^{p(p+1)/2}$.
- ③ If the normal distribution is non-singular, we have

$$\mathbb{E} \left[\frac{\partial \ln f(\mathbf{x}, \theta)}{\partial \theta} \left(\frac{\partial \ln f(\mathbf{x}, \theta)}{\partial \theta} \right)^\top \right] = -\mathbb{E} \left[\frac{\partial^2 \ln f(\mathbf{x}, \theta)}{\partial \theta \partial \theta^\top} \right].$$

Multivariate Cramer-Rao Inequality

Theorem 2

Under the regularity condition (everything is well-defined, integration and differentiation can be swapped), we have

$$N\mathbb{E}[(\mathbf{t} - \boldsymbol{\theta})(\mathbf{t} - \boldsymbol{\theta})^\top] \succeq \left(\mathbb{E} \left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right] \right)^{-1},$$

where $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$ and $f(\mathbf{x}, \boldsymbol{\theta})$ is the density of the distribution with respect to the components of $\boldsymbol{\theta}$.

- 1 Let $\mathbf{s} = \frac{\partial \ln g(\mathbf{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, where g is the density on N samples and $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$.
- 2 For unbiased estimator \mathbf{t} of $\boldsymbol{\theta}$, we have $\text{Cov}[\mathbf{t}, \mathbf{s}] = \mathbf{I}$.

Outline

- 1 Efficiency
- 2 Consistency**
- 3 Asymptotic Normality
- 4 Decision Theory
- 5 The Biased Estimator
- 6 Chi-Squared Distribution

Consistency

A sequence of vectors $\mathbf{t}_n = [t_{1n}, \dots, t_{pn}]^\top$ for $n = 1, 2, \dots$, is a consistent estimator of $\boldsymbol{\theta} = [\theta_1, \dots, \theta_p]^\top$ if

$$\text{plim}_{n \rightarrow \infty} t_{in} = \theta_i$$

for $i = 1, \dots, p$.

- 1 By the law of large numbers, the sample mean $\bar{\mathbf{x}}$ is a consistent estimator of $\boldsymbol{\mu}$ if the observations are i.i.d with mean $\boldsymbol{\mu}$ (normality is not involved).
- 2 The sample covariance matrix is also consistent since

$$\mathbf{S} = \frac{1}{N-1} \sum_{\alpha=1}^N (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top - \frac{N}{N-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top.$$

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality**
- 4 Decision Theory
- 5 The Biased Estimator
- 6 Chi-Squared Distribution

Asymptotic Normality

Let X_1, \dots, X_n be independent and identically distributed random variables with the same arbitrary distribution, mean μ , and variance σ^2 .

Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, then the random variable

$$Z = \lim_{n \rightarrow \infty} \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right)$$

is a standard normal distribution.

What about multivariate case?

Asymptotic Normality

Multivariate central limit theorem.

Theorem 3

Let p -component vectors $\mathbf{y}_1, \mathbf{y}_2, \dots$ be i.i.d with means $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\nu}$ and covariance matrices $\mathbb{E}[(\mathbf{y}_\alpha - \boldsymbol{\nu})(\mathbf{y}_\alpha - \boldsymbol{\nu})^\top] = \mathbf{T}$. Then the limiting distribution of

$$\frac{1}{\sqrt{n}} \sum_{\alpha=1}^n (\mathbf{y}_\alpha - \boldsymbol{\nu})$$

as $n \rightarrow +\infty$ is $\mathcal{N}(\mathbf{0}, \mathbf{T})$.

Characteristic Function and Probability

If \mathbf{x} does not have a density, the characteristic function uniquely defines the probability of any continuity interval.

Theorem 5

Let $\{F_j(\mathbf{x})\}$ be a sequence of cdfs, and let $\{\phi_j(\mathbf{t})\}$ be the sequence of corresponding characteristic functions. A necessary and sufficient condition for $F_j(\mathbf{x})$ to converge to a cdf $F(\mathbf{x})$ is that, for every \mathbf{t} , $\phi_j(\mathbf{t})$ converges to a limit $\phi(\mathbf{t})$ that is continuous at $\mathbf{t} = \mathbf{0}$. When this condition is satisfied, the limit $\phi(\mathbf{t})$ is identical with the characteristic function of the limiting distribution $F(\mathbf{x})$.

See the proof in Section 10.7 of “Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press”

Asymptotic Normality

Let

$$\mathbf{A}(n) = \sum_{\alpha=1}^N (\mathbf{x}_{\alpha} - \bar{\mathbf{x}}_N)(\mathbf{x}_{\alpha} - \bar{\mathbf{x}}_N)^{\top},$$

where $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independently distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $n = N - 1$. Then the limiting distribution of

$$\mathbf{B}(n) = \frac{1}{\sqrt{n}}(\mathbf{A}(n) - n\boldsymbol{\Sigma})$$

is normal with mean $\mathbf{0}$ and covariance $\mathbb{E}[b_{ij}(n)b_{kl}(n)] = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$.

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality
- 4 Decision Theory**
- 5 The Biased Estimator
- 6 Chi-Squared Distribution

Decision Theory

- 1 An observation random vector \mathbf{x} whose distribution P_θ depends on a parameter θ which is an element of a set Θ .
- 2 The statistician is to make a decision \mathbf{d} in a set \mathcal{D} .
- 3 A decision procedure is a function $\delta(\cdot)$ whose domain is the set of values of \mathbf{x} and whose range is \mathcal{D} .
- 4 The loss in making decision \mathbf{d} for the distribution of \mathbf{x} is a nonnegative function $L(\theta, \mathbf{d})$.
- 5 The evaluation of a procedure $\delta(\mathbf{x})$ is on the basis of the risk function

$$R(\theta, \delta) = \mathbb{E}_{\mathbf{x} \sim P_\theta} [L(\theta, \delta(\mathbf{x}))].$$

For example, the risk can be the mean squared error for univariate case

$$R(\theta, \delta) = \mathbb{E}_{\mathbf{x} \sim P_\theta} [(\delta(\mathbf{x}) - \theta)^2]$$

- ① A decision procedure $\delta(\mathbf{x})$ is as good as a procedure $\delta^*(\mathbf{x})$ if

$$R(\theta, \delta) \leq R(\theta, \delta^*),$$

and $\delta(\mathbf{x})$ is better than $\delta^*(\mathbf{x})$ if it holds with a strict inequality for at least one value of θ .

- ② A procedure $\delta^*(\mathbf{x})$ is inadmissible if there exists another procedure $\delta(\mathbf{x})$ that is better than $\delta^*(\mathbf{x})$.
- ③ A procedure is admissible if it is not inadmissible (i.e., if there is no procedure better than it) in terms of the given loss function.

Bayes Procedure

If the parameter θ can be assigned an a prior distribution, say, with density $\rho(\theta)$, then the average loss from use of a decision procedure $\delta(\mathbf{x})$ is

$$r(\rho, \delta) = \mathbb{E}_{\rho} [R(\theta, \delta)] = \mathbb{E}_{\theta \sim \rho} [\mathbb{E}_{\mathbf{x} \sim P_{\theta}} [L(\theta, \delta(\mathbf{x}))]] .$$

Given the a prior density ρ , the decision procedure $\delta(\mathbf{x})$ that minimizes $r(\rho, \delta)$ is the Bayes procedure, and the resulting minimum of $r(\rho, \delta)$ is the Bayes risk.

Bayes Procedure

If the density of \mathbf{x} given $\boldsymbol{\theta}$ is $f(\mathbf{x} | \boldsymbol{\theta})$, the joint density of \mathbf{x} and $\boldsymbol{\theta}$ is $f(\mathbf{x} | \boldsymbol{\theta})\rho(\boldsymbol{\theta})$ and the average risk of a procedure $\boldsymbol{\delta}(\mathbf{x})$ is

$$\begin{aligned} r(\rho, \boldsymbol{\delta}) &= \int_{\Theta} \int_{\mathcal{X}} L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) f(\mathbf{x} | \boldsymbol{\theta}) \rho(\boldsymbol{\theta}) d\mathbf{x} d\boldsymbol{\theta} \\ &= \int_{\mathcal{X}} \left(\int_{\Theta} L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) g(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta} \right) f(\mathbf{x}) d\mathbf{x}, \end{aligned} \quad (2)$$

where

$$f(\mathbf{x}) = \int_{\Theta} f(\mathbf{x} | \boldsymbol{\theta}) \rho(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad \text{and} \quad g(\boldsymbol{\theta} | \mathbf{x}) = \frac{f(\mathbf{x} | \boldsymbol{\theta}) \rho(\boldsymbol{\theta})}{f(\mathbf{x})}$$

are the marginal density of \mathbf{x} and the a posterior density of $\boldsymbol{\theta}$ given \mathbf{x} .

The procedure that minimizes $r(\rho, \boldsymbol{\delta})$ is one that for each \mathbf{x} minimizes the expression in braces on the right-hand side of (2), that is, the expectation of $L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x}))$ with respect to the a posterior distribution.

Bayes Procedure

If θ and δ are vectors and $L(\theta, \delta(\mathbf{x})) = (\theta - \delta(\mathbf{x}))^\top \mathbf{Q}(\theta - \delta(\mathbf{x}))$, where \mathbf{Q} is positive definite. Then we have

$$\begin{aligned}\mathbb{E}_{\theta|\mathbf{x}} [L(\theta, \delta(\mathbf{x}))] &= \mathbb{E}_{\theta|\mathbf{x}} [(\theta - \delta(\mathbf{x}))^\top \mathbf{Q}(\theta - \delta(\mathbf{x}))] \\&= \mathbb{E}_{\theta|\mathbf{x}} [(\theta - \mathbb{E}[\theta | \mathbf{x}])^\top \mathbf{Q}(\theta - \mathbb{E}[\theta | \mathbf{x}])] \\&\quad + \mathbb{E}_{\theta|\mathbf{x}} [(\theta - \mathbb{E}[\theta | \mathbf{x}])^\top \mathbf{Q}(\mathbb{E}[\theta | \mathbf{x}] - \delta(\mathbf{x}))] \\&\quad + \mathbb{E}_{\theta|\mathbf{x}} [(\mathbb{E}[\theta | \mathbf{x}] - \delta(\mathbf{x}))^\top \mathbf{Q}(\theta - \mathbb{E}[\theta | \mathbf{x}])] \\&\quad + \mathbb{E}_{\theta|\mathbf{x}} [(\mathbb{E}[\theta | \mathbf{x}] - \delta(\mathbf{x}))^\top \mathbf{Q}(\mathbb{E}[\theta | \mathbf{x}] - \delta(\mathbf{x}))] \\&= \mathbb{E}_{\theta|\mathbf{x}} [(\theta - \mathbb{E}[\theta | \mathbf{x}])^\top \mathbf{Q}(\theta - \mathbb{E}[\theta | \mathbf{x}])] \\&\quad + \mathbb{E}_{\theta|\mathbf{x}} [(\mathbb{E}[\theta | \mathbf{x}] - \delta(\mathbf{x}))^\top \mathbf{Q}(\mathbb{E}[\theta | \mathbf{x}] - \delta(\mathbf{x}))]\end{aligned}$$

and the minimum occurs at $\delta(\mathbf{x}) = \mathbb{E}[\theta | \mathbf{x}]$ the mean of the a posterior distribution.

Bayes Procedure

If $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independently distributed, each \mathbf{x}_α according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{\mu}$ has an a prior distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Phi})$, then the a posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \dots, \mathbf{x}_N$ is normal with mean

$$\boldsymbol{\Phi} \left(\boldsymbol{\Phi} + \frac{1}{N} \boldsymbol{\Sigma} \right)^{-1} \bar{\mathbf{x}} + \frac{1}{N} \boldsymbol{\Sigma} \left(\boldsymbol{\Phi} + \frac{1}{N} \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\nu} \quad (3)$$

and covariance matrix

$$\boldsymbol{\Phi} - \boldsymbol{\Phi} \left(\boldsymbol{\Phi} + \frac{1}{N} \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\Phi}.$$

If the loss function is

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) = (\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))$$

then the Bayes estimator of $\boldsymbol{\mu}$ is (3).

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality
- 4 Decision Theory
- 5 The Biased Estimator**
- 6 Chi-Squared Distribution

The Biased Estimator

The sample mean $\bar{\mathbf{x}}$ seems the natural estimator of the population mean $\boldsymbol{\mu}$ based on a sample from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

However, Stein (1956) showed $\bar{\mathbf{x}}$ is not admissible with respect to the mean squared loss when $p \geq 3$.

The Biased Estimator

Consider the loss function

$$L(\boldsymbol{\mu}, \mathbf{m}) = \|\boldsymbol{\mu} - \mathbf{m}\|_2^2,$$

where \mathbf{m} is an estimator of the mean $\boldsymbol{\mu}$.

If $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independently distributed to $\mathcal{N}_p(\boldsymbol{\mu}, N\mathbf{I})$, we have

$$\mathbb{E} \left[\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2 \right] = \sum_{\alpha=1}^p \text{Var}(\bar{x}_\alpha) = p.$$

The Biased Estimator

The estimator proposed by James and Stein is

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right) (\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}$$

where $\boldsymbol{\nu}$ is an arbitrary fixed vector and $p \geq 3$.

It holds that $\mathbb{E} \left[\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2 \right] < \mathbb{E} \left[\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2 \right]$.

For small values of $\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2$, the multiplier of $(\bar{\mathbf{x}} - \boldsymbol{\nu})$ is negative; that is, the estimator $\mathbf{m}(\bar{\mathbf{x}})$ is in the direction from $\boldsymbol{\nu}$ opposite to that of $\bar{\mathbf{x}}$.

The Biased Estimator

Table 3.2 gives values of the risk for $p = 10$ and $\sigma^2 = 1$. For example, if $\tau^2 = \|\boldsymbol{\mu} - \boldsymbol{\nu}\|^2$ is 5, the mean squared error of the James–Stein estimator is 8.86, compared to 10 for the natural estimator; this is the case if $\mu_i - \nu_i = 1/\sqrt{2} = 0.707$, $i = 1, \dots, 10$, for instance.

Table 3.2[†]. Average Mean Squared Error of the James–Stein Estimator for $p = 10$ and $\sigma^2 = 1$

$\tau^2 = \ \boldsymbol{\mu} - \boldsymbol{\nu}\ ^2$	$\mathcal{E}_{\boldsymbol{\mu}} \ m(Y) - \boldsymbol{\mu}\ ^2$
0.0	2.00
0.5	4.78
1.0	6.21
2.0	7.51
3.0	8.24
4.0	8.62
5.0	8.86
6.0	9.03

[†]From Efron and Morris (1977).

The Biased Estimator

The estimator proposed by James and Stein is

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right) (\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

For small values of $\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2$, the multiplier of $(\bar{\mathbf{x}} - \boldsymbol{\nu})$ is negative; that is, the estimator $\mathbf{m}(\bar{\mathbf{x}})$ is in the direction from $\boldsymbol{\nu}$ opposite to that of $\bar{\mathbf{x}}$.

We can improve $\mathbf{m}(\bar{\mathbf{x}})$ by using

$$\tilde{\mathbf{m}}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)^+ (\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu},$$

which holds that $\mathbb{E} \left[\|\tilde{\mathbf{m}}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2 \right] \leq \mathbb{E} \left[\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2 \right].$

Outline

- 1 Efficiency
- 2 Consistency
- 3 Asymptotic Normality
- 4 Decision Theory
- 5 The Biased Estimator
- 6 Chi-Squared Distribution**

Chi-Squared Distribution

If x_1, \dots, x_n are independent, standard normal random variables, then the sum of their squares,

$$y = \sum_{i=1}^n x_i^2,$$

is distributed according to the (central) chi-squared distribution (χ^2 -distribution) with n degrees of freedom.

We have $\mathbb{E}[y] = n$ and $\text{Var}[y] = 2n$.

Chi-Squared Distribution

The probability density function of the (central) chi-squared distribution is

$$f(y; n) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp\left(-\frac{y}{2}\right), & y > 0; \\ 0, & \text{otherwise,} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} \exp(-t) dt.$$

Chi-Squared Distribution

The derivation for the density is based on

- ① We have $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.
- ② For $y_1 = x^2$ with $x \sim \mathcal{N}(0, 1)$, the density function of y_1 is

$$\frac{1}{\sqrt{2\pi y_1}} \exp\left(-\frac{1}{2}y_1\right).$$

- ③ For beta function $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$, we have

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

- ④ If $F(z) = \int_{a(z)}^{b(z)} f(y, z) dy$, then

$$F'(z) = \int_{a(z)}^{b(z)} \frac{\partial f(y, z)}{\partial z} dx + f(b(z), z)b'(z) - f(a(z), z)a'(z).$$

Noncentral Chi-Squared Distribution

If x_1, \dots, x_n are independent and each x_i are normally distributed random variables with means μ_i and unit variances, then the sum of their squares,

$$y = \sum_{i=1}^n x_i^2,$$

is distributed according to the noncentral Chi-squared distribution with n degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^n \mu_i^2.$$

We have $\mathbb{E}[y] = n + \lambda$ and $\text{Var}[y] = 2n + 4\lambda$.

Noncentral Chi-Squared Distribution

If y_1, \dots, y_k are independent and each y_i is distributed according to the noncentral chi-squared distribution with n_i degrees of freedom and noncentrality parameter λ_i , then

$$\sum_{i=1}^k y_i \sim \chi_{n_1 + \dots + n_k}^2 \left(\sum_{i=1}^k \lambda_i \right).$$

Theorem 4

If the n -component vector \mathbf{y} is distributed according to $\mathcal{N}(\boldsymbol{\nu}, \mathbf{T})$ with $\mathbf{T} \succ \mathbf{0}$, then

$$\mathbf{y}^\top \mathbf{T}^{-1} \mathbf{y} \sim \chi_n^2 \left(\boldsymbol{\nu}^\top \mathbf{T}^{-1} \boldsymbol{\nu} \right).$$

If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is the central χ^2 -distribution.