
高等数理统计讲义

Advanced Mathematical Statistics Lecture Notes

马学俊

献给我的家人、恩师和所有在学术道路上帮助我的人

马学俊, 副教授, 苏州大学数学科学学院统计系, 主要从事海量数据分析、高维数据分析、统计计算、非参数回归等统计模型及其应用等研究。个人主页 <https://xuejunma.github.io>.
We would like to thank Professor [Tian Maozai \(RUC\)](#) and Professor [Wang Wanjie \(NUS\)](#) for sending slides and beneficial discussions.

目录

1	Review of Basic Probability	1
1.1	Outline	1
1.2	Random variable	2
1.3	Expection	18
1.4	Differentiating Under An Integral Sign	29
1.5	Important Distribution	32
1.6	Statistics	34
1.7	Moment inequalities	35
	第 1 讲 练习	42
2	Convergence of Random Variables	47
2.1	Look into Sample mean	47
2.2	Convergence in L_r	52

2.3	Relationship	65
2.4	Properties of Convergence	68
2.5	Stochastic Orders	70
	第 2 讲 练习	72
3	Law of Large Numbers and Central Limit Theorem	76
3.1	Law of Large Numbers	77
3.2	Central Limit Theorem (CLT)	84
	第 3 讲 练习	87
4	Edgeworth Expansion and Models	89
4.1	The Delta Method	89
4.2	The Edgeworth Expansion	93
	第 4 讲 练习	97
5	Principle of Data Reduction	99
5.1	Introduction	100

5.2	Population and Sample	100
5.3	Parameter Estimation	102
5.4	Some popular models	106
5.5	Statistics	109
6	Point estimation	137
6.1	Estimators	137
6.2	Method of Moments(MoM)	138
6.3	Maximum Likelihood Estimator(MLE)	141
6.4	Bayes Estimators	147
6.5	EM Algorithm	158
6.6	Methods of Evaluating Estimators	163
	第 6 讲 练习	184
7	Asymptotic properties of estimators	185
7.1	Consistency	185
7.2	Consistency of MLE	188

7.3	Asymptotic Properties	192
7.4	Asymptotic Relative Efficiency	200
7.5	Robustness	201
8	Hypothesis Testing	211
8.1	Evaluation of a test	213
8.2	Evaluation	215
8.3	Generally used tests	217
8.4	The Wald Test	221
8.5	The Likelihood Ratio Test (LRT)	223
8.6	p -values	228
8.7	The Permutation Test	230
8.8	Multiple Testing	232
9	Confidence Sets	242
9.1	Confidence Sets	242

第 1 讲 Review of Basic Probability

内容提要

- | | |
|---|--|
| <input type="checkbox"/> Sample space; Measure; Random variable | <input type="checkbox"/> Moment Generating Function |
| <input type="checkbox"/> Transformation | <input type="checkbox"/> Characteristic function |
| <input type="checkbox"/> Independence | <input type="checkbox"/> Common distributions |
| <input type="checkbox"/> Expectation | <input type="checkbox"/> Sample mean; Sample variance; Sample distribution |
| <input type="checkbox"/> Conditional expectation | <input type="checkbox"/> Moment inequalities |

1.1 Outline

- Brief review of basic probability and statistics
 - Why is a random variable?

- Transformations; independence; expectation
- Important distributions
- Some statistics

1.2 Random variable

1.2.1 Sample space and Measure

What do we mean by randomness?

- We construct an experiment, yet the result of the experiment has many possibilities.
 - Flip a coin, the result can be either head or tail
- Although we can not know the result beforehand, we do have some information about the result.
 - Approximately, there is equal chance for a head and a tail
- Randomness: the **uncertainty** of experiment results

Question: How to describe our **information**?

- Information 1. Possible outcomes

定义 1.1 (Sample space (Outcome space))

Let Ω be a sample space, which is a set containing all possible outcomes.



- Information 2. Probabilities for these possible outcomes
 - σ -field \mathcal{F} : a set of subsets of Ω which satisfies 3 rules.
 - Measurable space: (Ω, \mathcal{F})
 - Event(measurable sets): element of \mathcal{F}
 - Probability measure P : for any element in the σ -field, assign it a probability, indicating the chance this event will happen
- $(\Omega; \mathcal{F}; P)$ (Probability space, measure space) is our information about the possible outcomes of this experiment. In short, we write it as the sample space Ω with probability P , or just Ω if there is no confusion.

定义 1.2 (σ -field)

Let \mathcal{F} be a collection of subsets of a sample space. \mathcal{F} is called a σ -field (or σ -algebra) if and only if it has the following properties.

- The empty set $\phi \in \mathcal{F}$.
- If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$.
- If $A_i \in \mathcal{F}, i = 1, 2, \dots$, then their union $\cup A_i \in \mathcal{F}$.



- Measurable space: (Ω, \mathcal{F})
- Event (measurable sets): element of \mathcal{F}
- $\sigma(A) = \{\phi, A, A^c, \Omega\}$.
- Flip a coin, the result can be either head or tail $\Omega = \{H, T\}, \mathcal{F} = \{\dots\}$

定义 1.3 (Measure)

Let Measurable space (Ω, \mathcal{F}) , ν be a measurable space. A set function ν defined on \mathcal{F} is called a measure if and only if it has the following properties.

- $0 \leq \nu(A) \leq \infty$, for any $A \in \mathcal{F}$
- $\nu(\phi) = 0$
- If $A_i \in \mathcal{F}, i = 1, 2, \dots$, and A_i 's are disjoint, i.e. $A_i \cap A_j = \emptyset$ for any $i \neq j$, then $\nu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \nu(A_i)$



- measure space: $(\Omega, \mathcal{F}, \nu)$
- probability measure $\nu(\Omega) = 1$. We usually denote it by P instead of ν , (Ω, \mathcal{F}, P) .
- Flip a coin, the result can be either head or tail $\Omega = \{H, T\}, \mathcal{F} = \{\dots\}$
 - $\nu(A) = |A|$ the number of elements in $A \in (\mathcal{F})$.
 - $P(A) = \frac{|A|}{|\Omega|}$

1.2.2 Random Variables

What is of interest?

- Manufacturers Ω : all the combinations of good light bulbs and defective light bulbs. Need: proportion of defective light bulbs from a lot
- Market researchers Ω : survey results of all consumers for one product. Need: preference of all consumers about this product, with a scale 1-10.

Our interest:

- Not the details of Ω , but a special measurable characteristic of the outcomes!
- A random variable, is a mapping from Ω to R , which draws the measurable characteristic of interest

Example: an opinion poll. 50 people; 1: agree; 0 disagree:

- Ω has 2^{50} elements.
- interest: the number of people who agree out of 50. X = number of 1s recorded out of 50.
 $\mathbb{X} = \{0, 1, 2, \dots, 50\}$

定义 1.4 (Random Variable)

- Let (Ω, \mathcal{F}) and $(\mathcal{R}, \mathcal{B})$ (\mathcal{B} : Borel σ -field) be measurable spaces
- X is a function from Ω to \mathcal{R} . The function X is called a **random variable** (r.v.; measurable function) if and only if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \subset \mathcal{F}$$

for any $B \in \mathcal{B}$.

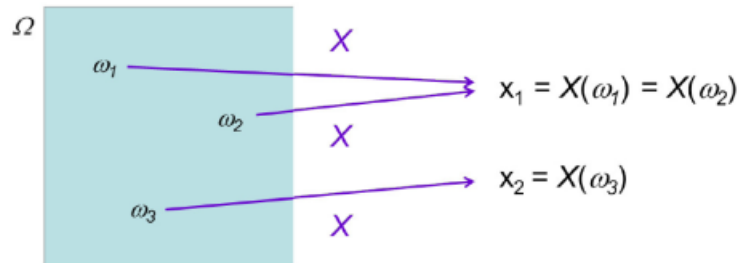


- Suppose we have a sample space

$$\Omega = \{\omega_1, \dots, \omega_n\}$$

with a probability function P .

- We defined a random variable X with range $\mathbb{X} = \{x_1, \dots, x_m\}$.



- We write

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

$$P_X(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})$$

where P_X is an **induced probability** function χ .

- Notations:
 - Upper-case letters X, Y, Z, \dots to denote r.v.'s

- Lower-case letters x, y, z, \dots to denote their possible values.

例题 1.1 (Example 1.4.3)

- Consider the experiment of tossing a coin three times.
- H :Head; T :Tail.
- X :the number of heads obtained in the three tosses.

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

- $\mathbb{X} = \{0, 1, 2, 3\}$. The induced probability function on \mathcal{X} is given by

x	0	1	2	3
$P_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$P_X(X = 1) = P(\{HTT, THT, TTH\}) = \frac{3}{8}$$

定义 1.5 ((Definition 1.5.1 Cumulative Density Function))

The cumulative distribution function (CDF) of a random variable is defined by

$$F(x) = P(X \leq x); -\infty < x < \infty$$



For all CDF's; there is

- $F(x)$ is right-continuous. At each x , $\lim_{n \rightarrow \infty} F(y_n) = F(x)$ for any sequence $y_n \rightarrow x$ with $y_n > x$.
- $F(x)$ is non-decreasing.
- $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$.

Any $F(x)$ satisfying Properties 1-3 is a CDF for some random variable.

例题 1.2(Example 1.5.5)

$$F_X(x) = \frac{1}{1 + e^{-x}}$$

- $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$

-

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1 + e^{-x})^2} > 0$$

- If X is discrete, then its **probability mass function (pmf)** is

$$p_X(x) = p(x) = P(X = x)$$

- If X is continuous, then its **probability density function (pdf)** satisfies

$$P(X \in A) = \int_A f_X(x) dx = \int_A f(x) dx = \int_A dF(x)$$

and $f_X(x) = f(x) = F'(x)$.

- We say that X and Y have the **same distribution** (i.e. $X \stackrel{D}{=} Y$) if $P(X \in A) = P(Y \in A)$ for all A . $X \stackrel{D}{=} Y$ if and only if $F_X(t) = F_Y(t)$

1.2.3 Transformation

Given a r.v. X with density function $f_X(\cdot)$, it is often that we are interested in a **transformation** $Y = g(X)$ which is defined as a known function g (either **one-to-to** or **many-to-one**) of X .

- Obviously, the composite function $g \circ X$ defines a new r.v. Y from Ω to R .

- Let $Y = g(X)$.

$$\begin{aligned} P(Y \in A) &= P(g(X) \in A) \\ &= P(X \in g^{-1}(A)) \end{aligned} \tag{1.1}$$

where $g^{-1}(A) = \{x \in R, g(x) \in A\}$. In particular,

$$F_Y(y) = Pr\{Y \leq y\} = P(X \in g^{-1}(-\infty, y])$$

If X has pdf $f_X(x)$, then

$$F_Y(y) = \int_{g^{-1}(-\infty, y]} f_X(x) dx = \int_{\{x: g(x) \leq y\}} f_X(x) dx$$

例题 1.3 (Example 2.1.2) Suppose X has a uniform distribution on the interval $(0, 2\pi)$, that is

$$f_X(x) = \begin{cases} 1/2\pi, & 0 < x < 2\pi, \\ 0, & \text{otherwise} \end{cases} \tag{1.2}$$

Consider $Y = \sin^2(X)$

$$\begin{aligned} P(Y \leq y) &= P(X \leq x_1) + P(x_2 \leq X \leq x_3) + P(X \geq x_4) \\ &= 2P(X \leq x_1) + 2P(x_2 \leq X \leq \pi) \end{aligned} \tag{1.3}$$

- If g is increasing,

$$F_Y(y) = F_X(g^{-1}(y))$$

.

- If g is decreasing,

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

.

定理 1.1 (Theorem 2.1.5)

Let X have probability distribution function(pdf) $f_X(x)$ and $Y = g(X)$, where g is a monotone function. Let

$$\mathcal{Y} = \{y : g^{-1}(y) \text{ is a possible value of } X\}$$

. Suppose $f_X(x)$ is continuous and that $g^{-1}(y)$ has a continuous derivative on \mathcal{Y} . Then the pdf

on Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise} \end{cases} \quad (1.4)$$



例题 1.4(Example 2.1.4) $X \sim f_X(x) = 1I(0 < x < 1)$, $F_X(x) = x$. $Y = g(x) = -\log x$, find its distribution.

Proof:

- $Y = g(x) = -\log x \Rightarrow x = e^{-y}, g^{-1}(y) = e^{-y}$
- g is a decreasing function.

$$\frac{d}{dx} g(x) = \frac{d}{dx} (-\log x) = -\frac{1}{x} < 0, 0 < x < 1$$

•

$$\begin{aligned}
F_Y(y) &= P_Y(Y \leq y) = P_X(g(x) \leq y) \\
&= P_X(X \geq g^{-1}(y)) \\
&= 1 - P_X(X \leq g^{-1}(y)) \\
&= 1 - e^{-y}
\end{aligned}$$

例题 1.5(Example 2.1.6) Let

$$f_X(x) = \frac{1}{(n-1)!\beta^n} x^{n-1} e^{-x/\beta}, 0 < x < \infty$$

be the Gamma pdf $Y = 1/X$. Find the pdf of Y

Proof. $g^{-1}(y) = 1/y$, $\mathcal{Y} = (0, \infty)$, $\left| \frac{d}{dy} g^{-1}(y) \right| = 1/y^2$. Therefore for all $y > 0$,

$$\begin{aligned}
f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\
&= \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y} \right)^{n-1} e^{-1/(\beta y)} \frac{1}{y^2} \\
&= \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y} \right)^{n+1} e^{-1/(\beta y)}
\end{aligned} \tag{1.5}$$

- A special case of a pdf known as the inverted Gamma distribution.

定理 1.2 (Theorem 2.1.8)

Let X have pdf $f_X(x)$, let $Y = g(X)$. Suppose there exists a partition A_0, A_1, \dots, A_k such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i .

$$P(X \in \cup_{i=1}^k A_i) = 1.$$

Further, we have $g(\cdot)$ is monotone if restricted to $A_i, i = 1, 2, \dots, k$. Let

$$g_i^{-1}(y) = \{x \in A_i : g(x) = y\}$$

and assume $g_i^{-1}(y)$ has continuous derivative on \mathcal{Y} for each i . Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^k f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise} \end{cases}$$



- **Remark** Unfortunately, we found the above Theorem has very little practical use.

例题 1.6 (Example 2.1.9) Let $X \sim N(0, 1)$, $Y = X^2$, we may use the above theorem to find the pdf of Y .

Proof:

- $g(x) = x^2$ is monotone on $(-\infty, 0)$ and on $(0, \infty)$.

- $\mathcal{Y} = (0, \infty)$.

$$A_0 = \{0\}$$

$$A_1 = (-\infty, 0), g_1(x) = x^2, g_1^{-1}(y) = -\sqrt{y}$$

$$A_2 = (0, \infty), g_2(x) = x^2, g_2^{-1}(y) = \sqrt{y}$$

The pdf Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \Phi(\sqrt{y}) \frac{1}{2} \frac{1}{\sqrt{y}} + \Phi(-\sqrt{y}) \frac{1}{2} \frac{1}{\sqrt{y}} = \frac{1}{\sqrt{y}} \phi(\sqrt{y})$$

定理 1.3 (Theorem 2.1.10 Probability integral transform)

Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is

$$P(Y \leq y) = y, 0 < y < 1.$$



- $F_X^{-1}(\tau) = \inf\{x : F_X(x) \geq \tau\}$

- Proof:

$$\begin{aligned}
 P_Y(Y \leq y) &= P_X(F_X(x) \leq y) \\
 &= P_X(F_X^{-1}[F_X(x)] \leq F_X^{-1}(y)) \\
 &= P_X(X \leq F_X^{-1}(y)) \\
 &= F_X(F_X^{-1}(y)) \\
 &= y
 \end{aligned}$$

定理 1.4 (Theorem 4.2.10)

Two r.v.'s X and Y are **independent** if and only if

$$P(X \in A; Y \in B) = P(X \in A)P(Y \in B)$$

for all A and B .



- $F(x, y) = F(x)F(y)$ for any x and y , $f(x, y) = f(x)f(y)$ or $p(x, y) = p(x)p(y)$
- When X and Y are independent, $h(X)$ and $g(Y)$ are also independent, if h and g are well-defined functions.

1.3 Expectation

- Definition:

$$E(X) = \sum_x xp(x)$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- Properties:

- $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$

- $E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$

- If X_1, \dots, X_n are independent, then

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i)$$

- Example 2.2.2

$X \sim \exp(\lambda),$

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda} x > 0.$$

Then $E[X] = \lambda.$

- [Example 2.2.3](#)

$X \sim \text{Binomial}(n,p),$

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \dots.$$

Then $E[X] = np.$

- [Example 2.2.4](#)

$X \sim \text{Cauchy},$

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2} \quad -\infty < x < \infty.$$

Then $E[X]$ is not defined!(or do not exist).

- Mixed normal distribution

$$X = 0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$$

Mixed normal distribution

```

1 rm(list=ls())
2 n <- 1000
3 comp <- sample(c(0, 1), size = n, prob = c(0.5, 0.5), replace = T)
4 x <- rnorm(n, mean = ifelse(comp == 0, -1, 1), sd = ifelse(comp == 0, 0.5, 0.5))
5 plot(density(x), main="")

```

定理 1.5 (Theorem 2.2.5)

Let X be a r.v. and let a, b and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist.

(1) $E[ag_1(X) + bg_2(X) + c] = aE[g_1(X)] + bE[g_2(X)] + c$

(2) If $g_1(x) \geq 0$ for all x , then $E[g_1(X)] \geq 0$.

(3) If $g_1(x) \geq g_2(x)$ for all x , then $E[g_1(X)] \geq E[g_2(X)]$

(4) If $a \leq g_1(x) \leq b$ for all x , then $a \leq E[g_1(X)] \leq b$



例题 1.7(Example 2.2.6) $E(X)$ is the "center" of a distribution (or its r.v.) in the sense that

$$\min_b E(X - b)^2 = E[X - EX]^2.$$

- **Homework**

$$\min_b E \rho_\tau(X - b)$$

Remark: $\rho_\tau(t) = \tau t I(t \geq 0) + (\tau - 1)t I(t \leq 0)$.

1.3.1 Variance & Standard Deviation

- Motivation: Describe the "spread" of r.v.
- Definition. $Var(x) = E[(x - \mu)^2]$, where $\mu = E(X)$, $sd(X) = \sqrt{Var(x)}$.
- Properties.

- $Var(X) = E(X^2) - [E(X)]^2$

- If X_1, \dots, X_n are independent, then

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i)$$

- The covariance is

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

and the **correlation coefficient** is

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

- For any two r.v.s with variance existed,

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

1.3.2 Conditional Expectation

- Conditional Expectation of X when Y is given as y is that

- $E(X|Y = y) = \sum_x x p_{X|Y}(x|y)$ for discrete r.v.

- $\mathbf{E}(X|Y = y) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)dx$ for cont. r.v.
- Interpretation: Note that $X|Y = y$ is a new r.v., $\mathbf{E}(X|Y = y)$ is the expectation on this r.v.
- Law of Total Expectation

$$\mathbf{E}[\mathbf{E}(X|Y)] = \mathbf{E}(X)$$

- Law of Total Variance

$$\text{Var}(X) = \text{Var}[\mathbf{E}(X|Y)] + \mathbf{E}[\text{Var}(X|Y)]$$

定理 1.6 (Theorem 4.4.3)

If X and Y are any two r.v.s, then

$$\mathbf{E}(X) = \mathbf{E}[\mathbf{E}(X|Y)]$$



Proof:

$$\begin{aligned}
 \mathbf{E}X &= \int \int x f(x, y) dx dy \\
 &= \int \left[\int x f(x|y) dx \right] f_Y(y) dy \\
 &= \int \mathbf{E}(X|y) f_Y(y) dy = \mathbf{E}[\mathbf{E}(X|Y)]
 \end{aligned}$$

In general, the conditional expectation $\mathbf{E}[X|Y]$ can be defined as a r.v. $g(Y)$ such that

$$\mathbf{E}[(X - g(Y))^2] = \inf_{\text{among all reasonable function } h} \mathbf{E}[(X - h(Y))^2]$$

or $\mathbf{E}[X|Y]$ is the function of Y which is "closest" to X in terms of mean square error.

例题 1.8(9Example 4.4.1) $Y \sim$ Number of eggs lay by a mother fish, and $X \sim$ Number of survivors (young fish). On the average, how many eggs will survive?

Then it is reasonable to assume

$$Y \sim \text{Poisson}(\lambda)$$

$$X|Y \sim \text{Binomial}(Y, p)$$

So,

$$\begin{aligned}EX &= E[E(X|Y)] \\&= E(pY) \\&= p\lambda\end{aligned}$$

例题 1.9(Example 4.4.5)

$$X|Y \sim \text{Binomial}(Y, p)$$

$$Y|\Lambda \sim \text{Poisson}(\Lambda)$$

$$\Lambda \sim \text{exponential}(\beta)$$

Proof:

$$\begin{aligned}E[X] &= E[E(X|Y)] \\&= pE(Y) \\&= pE[E(Y|\Lambda)] \\&= pE[\Lambda] \\&= p\beta.\end{aligned}$$

定理 1.7 (Theorem 4.4.7)

For any two random variables X and Y

$$\text{Var}(X) = \text{Var}[\mathbf{E}(X|Y)] + \mathbf{E}[\text{Var}(X|Y)]$$

provided that the expectation exists.



Proof:

$$\begin{aligned} \text{Var}(X) &= \mathbf{E}\left\{[X - \mathbf{E}(X|Y) + \mathbf{E}(X|Y) - \mathbf{E}X]^2\right\} \\ &= \mathbf{E}\left\{[X - \mathbf{E}(X|Y)]^2 + [\mathbf{E}(X|Y) - \mathbf{E}X]^2\right. \\ &\quad \left.+ 2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X]\right\} \\ &= \mathbf{E}\{[X - \mathbf{E}(X|Y)]^2\} + \mathbf{E}\{[\mathbf{E}(X|Y) - \mathbf{E}X]^2\} \\ &= \mathbf{E}[\text{Var}(X|Y)] + \text{Var}[\mathbf{E}(X|Y)] \end{aligned}$$

$$\mathbf{E}\left\{2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X]\right\} = \mathbf{E}[\mathbf{E}(Z|Y)]$$

1.3.3 Moment Generating Function and Characteristic Function

- Moment Generating Function(MGF)
 - Definition: $M_X(t) = E(e^{tX})$: a function of t , not r.v.
 - If $Y = aX + b$, $M_Y(t) = e^{bt} M_X(at)$
 - If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$
- Characteristic Function
 - Definition: $\phi_X(t) = E[e^{itX}]$: a function of t ; $i = \sqrt{-1}$.
 - Bounded: $|\phi(t)| \leq 1$
 - If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

An example of two distribution functions but with the same moments.

例题 1.10 (Example 2.3.10) Consider the two pdfs given by

$$f_1(x) = \frac{1}{\sqrt{2\pi x}} e^{-(\log x)^2/2}, 0 \leq x < \infty,$$

$$f_2(x) = f_1(x)[1 + \sin(2\pi \log x)], 0 \leq x < \infty.$$

Then it can be shown if $X_1 \sim f_1(x)$,

$$E[X_1^r] = e^{r^2/2}, r = 0, 1, \dots$$

Now suppose that $X_2 \sim f_2(x)$, we have for $r = 0, 1, \dots$

$$\begin{aligned} E[X_2^r] &= \int_0^\infty x^r f_1(x) [1 + \sin(2\pi \log x)] dx \\ &= E[X_1^r] + \int_0^\infty x^r f_1(x) \sin(2\pi \log x) dx \\ &= \int_0^\infty x^r \frac{1}{\sqrt{2\pi x}} 2e^{-(\log x)^2} \sin(2\pi \log x) dx \quad y = \log x - r \\ &= \int_{-\infty}^\infty e^{(y+r)r} \frac{1}{\sqrt{2\pi}} e^{-(y+r)^2/2} \sin(2\pi(y+r)) dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(y^2-r^2)} \sin(2\pi y) dy \cdot \cos(2\pi r) \\ &\quad + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(y^2-r^2)} \cos(2\pi y) dy \cdot \sin(2\pi r) \\ &= 0 \quad r = 0, 1, \dots \end{aligned}$$

Since $e^{-\frac{1}{2}(y^2-r^2)} \sin(2\pi y)$ is an odd function.

However, we have the following theorem.

定理 1.8 (Theorem 2.3.11)

Let $F_X(x)$ and $F_Y(y)$ be two CDFs all of whose moments exist.

1. If F_X and F_Y have bounded support, then $F_X(u) = F_Y(u)$ for all u iff $EX^r = EY^r$ for all $r = 0, 1, 2, \dots$
2. If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u .



1.4 Differentiating Under An Integral Sign

If a, b are finite and $f(x, \theta)$ is differential with respect to θ . Then we have

$$\frac{d}{d\theta} \int_a^b f(x, \theta) dx = \int_a^b \frac{\partial}{\partial \theta} f(x, \theta) dx$$

But in statistics, we often need to evaluate $\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx$, which may or may not be $\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx$.

定理 1.9 (Theorem 2.4.2)

Suppose the function $h(x, y)$ is continuous at y_0 for each x , and there exists a function $g(x)$ satisfying

a) $|h(x, y)| \leq g(x)$, for all x and y ;

b) $\int_{-\infty}^{\infty} g(x) dx < \infty$.

Then

$$\lim_{y \rightarrow y_0} \int_{-\infty}^{\infty} h(x, y) dx = \int_{-\infty}^{\infty} \lim_{y \rightarrow y_0} h(x, y) dx$$

.



Apply the above Theorem to the differentiation case, then we have

- **Theorem 2.4.3** Suppose $f(x, \theta)$ is differentiable at $\theta = \theta_0$, and there exists a function $g(x, \theta_0)$ and a constant $\delta > 0$ such that

a) $\left| \frac{f(x, \theta_0 + \Delta) - f(x, \theta_0)}{\Delta} \right| \leq g(x, \theta_0)$, for all x and $|\Delta| \leq \delta$; b) $\int_{-\infty}^{\infty} g(x, \theta_0) dx < \infty$. Then

$$\left. \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx \right|_{\theta=\theta_0} = \int_{-\infty}^{\infty} \left[\left. \frac{\partial}{\partial \theta} f(x, \theta) \right|_{\theta=\theta_0} \right] dx \quad (*)$$

- **Corollary** Suppose that there exists $\delta > 0$ and function $g(x, \theta)$ such that $|\frac{\partial}{\partial \theta} f(x, \theta)|_{\theta=\theta'} \leq g(x, \theta)$, for all θ' with $|\theta' - \theta| < \delta$, and $\int_{-\infty}^{\infty} g(x, \theta) dx < \infty$. Then (*) holds.
- **Remark** Finding bound $g(x, \theta)$ is cumbersome. We need to know that differentiating under the integral sign is not always automatic. In most situations, we just do it!!
- **Example 2.4.6** $X \sim N(\mu, 1)$,

$$M_X(t) = E(e^{tX}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2} dx,$$

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

For the exchange of operation of **differentiation** and **summation**, we have

- **Theorem 2.4.8** Suppose that the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges for all θ in an interval (a, b) and
 1. $\frac{\partial}{\partial \theta} h(\theta, x)$ is continuous in θ for each x ;
 2. $\sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x)$ converges uniformly on every closed bounded subinterval of (a, b) .

Then

$$\frac{d}{d\theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x)$$

- **Theorem 2.4.10** Suppose that the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges uniformly on $[a, b]$ and that, for each x , $h(\theta, x)$ is a continuous function of θ . Then

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x) d\theta = \sum_{x=0}^{\infty} \int_a^b h(\theta, x) d\theta$$

1.5 Important Distribution

- Discrete distributions:
 - Bernoulli r.v.: $X \sim \text{Bernoulli}(p)$, $p(1) = p$, $p(0) = 1 - p$, $p(x) = 0$ if $x \neq 0$ and $x \neq 1$. It can be written as $p^x(1 - p)^{1-x}$ for $x = 0, 1$.
 - Binomial r.v.: $X \sim \text{Binomial}(n, p)$, $p(x) = \binom{n}{x} p^x q^{n-x}$, $x = 0, 1, 2, \dots, n$. Summation of n Bernoulli random variables.

- Poisson r.v.: $X \sim Pois(\lambda), p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, 2, \dots, n.$
- Continuous distribution
 - Uniform r.v.: $X \sim Unif(a, b), f(x) = \frac{1}{b-a}, x \in (a, b)$
 - Exponential r.v.: $X \sim Exp(\lambda), f(x) = \lambda e^{-\lambda x}$
 - Normal r.v.: $X \sim N(\mu, \sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- The d random vector $X \sim N(\mu, \Sigma),$

$$f(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

- $AX + b \sim N(A\mu + b, A\Sigma^{-1}A^T)$
- Conditional distribution.

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

then

$$X_1|X_2 \sim N(\mu_1 + \Sigma_{11}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

1.6 Statistics

- Sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$
- Sample variance: $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
- Sampling distribution of \bar{X}_n : $G_n(t) = P(\bar{X}_n \leq t)$

When it is normal:

- If $X \sim N(\mu, \Sigma^2)$, then \bar{X}_n and S_n^2 are independent,

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$
$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

1.7 Moment inequalities

定理 1.10 (Lemma 4.7.1)

Let a and b be any two positive numbers, and let p and q be any positive numbers satisfying

$$\frac{1}{p} + \frac{1}{q} = 1$$

. Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality holds if and only if $a^p = b^q$.



- **Proof:** Consider for fixed b (or a),

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

with equality if and only if $a^p = b^q$.

定理 1.11 (Theorem 4.7.2(Holder's Inequality))

Let X and Y be any two random variables. Let p and q be any positive numbers satisfying

$$\frac{1}{p} + \frac{1}{q} = 1$$

Then

$$|E(XY)| \leq E|XY| \leq (E|X|^p)^{\frac{1}{p}} (E|Y|^q)^{\frac{1}{q}}$$



Proof: In the inequality (1), let

$$a = \frac{|X|}{(E|X|^p)^{\frac{1}{p}}}, b = \frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}$$

After some simplification, and take expectation on the two sides of the inequality. The result can be obtained.

- **Theorem 4.7.3(Cauchy-Schwarz Inequality)**

For any two random variables X and Y ,

$$|E(XY)| \leq E|XY| \leq (E|X|^2)^{\frac{1}{2}} (E|Y|^2)^{\frac{1}{2}}$$

- **Example 4.7.4 (Covariance Inequality)**

If X and Y have means μ_X and μ_Y , and variances σ_X^2 and σ_Y^2 , respectively. We can apply the Cauchy-Schwarz Inequality to get

$$(\text{Cov}(X, Y))^2 \leq \sigma_X^2 \sigma_Y^2$$

- **Example**

Let $p > 1$, then apply Holder's Inequality. For any random variables X ,

$$E|X| \leq \{E|X|^p\}^{\frac{1}{p}}$$

If $1 < r < s$, we have (Liapounov's Inequality)

$$(E|X|^r)^{\frac{1}{r}} \leq (E|X|^s)^{\frac{1}{s}}$$

- **Proof** Let q be such that $\frac{1}{p} + \frac{1}{q} = 1$, then

$$E|X| = E|X| \cdot 1 \leq (E|X|^p)^{\frac{1}{p}} (E1^q)^{\frac{1}{q}} = (E|X|^p)^{\frac{1}{p}}$$

.

- **Proof** Let s be such that $s = pr$, then $s > 1$.

$$E(|X|^r) \leq (E|X|^s)^{\frac{1}{p}}$$

.

定理 1.12 (Theorem 4.7.5(Minkowski's Inequality))

Let X and Y be any two random variables. Then for $1 < p < \infty$

$$[E|X + Y|^p]^{\frac{1}{p}} \leq (E|X|^p)^{\frac{1}{p}} + (E|Y|^p)^{\frac{1}{p}}$$



Proof:

$$\begin{aligned} E|X + Y|^p &= E(|X + Y||X + Y|^{p-1}) \\ &\leq E(|X||X + Y|^{p-1}) + E(|Y||X + Y|^{p-1}) \end{aligned} \quad (1.6)$$

Using Holder's Inequality,

$$E(|X||X + Y|^{p-1}) \leq (E|X|^p)^{\frac{1}{p}} [E|X + Y|^{q(p-1)}]^{\frac{1}{q}} \quad (1.7)$$

where q is such that $\frac{1}{p} + \frac{1}{q} = 1$ or $\frac{1}{q} = 1 - \frac{1}{p}$, i.e. $q = \frac{p}{p-1}$ or $q(p-1) = p$. Similarly,

$$E(|Y||X + Y|^{p-1}) \leq (E|Y|^p)^{\frac{1}{p}} [E|X + Y|^{q(p-1)}]^{\frac{1}{q}} \quad (1.8)$$

So combine (1.6) and (1.7) with (1.8), divide through by $[E(|X + Y|^{q(p-1)})]^{1/q}$, we have

$$E|X + Y|^p \leq (E|X + Y|^p)^{\frac{p-1}{p}} [(E|X|^p)^{\frac{1}{p}} + (E|Y|^p)^{\frac{1}{p}}]$$

定理 1.13

For any random variable X , if $g(x)$ is a convex function, then

$$Eg(X) \geq g(EX)$$

- Equality holds if and only if, for any line $a + bx$ that is tangent to $g(x)$ at $x = EX$, $P(g(X) = a + bX) = 1$.
- If $g(x)$ is linear, $g(EX) = a + bEX = Eg(X)$.



Remark For any twice differentiable function $g(x)$, it is convex if $g''(x) \geq 0$ for all x .

例题 1.11 An inequality for means Let a_1, a_2, \dots, a_n be n non-negative numbers. Define

$$a_A = \frac{1}{n}(a_1 + a_2 + \dots + a_n)$$

$$a_G = [a_1 a_2 \dots a_n]^{1/n} \quad a_H = \frac{1}{\frac{1}{n}(\frac{1}{a_1} + \dots + \frac{1}{a_n})}$$

An inequality relating to these means is

$$a_H \leq a_G \leq a_A$$

Remark The above inequality gives a reason for Maximum Likelihood Estimation(MLE).

Proof: Let X be a random variable with range a_1, \dots, a_n , and $P(X = a_i) = 1/n, n = 1, \dots, n$.

Since $\log x$ is a concave function, $E \log X \leq \log(EX)$, hence

$$\begin{aligned} \log a_G &= \frac{1}{n} \sum_{i=1}^n \log a_i = E \log X \leq \log(EX) \\ &= \log \left(\frac{1}{n} \sum_{i=1}^n a_i \right) = \log a_A \end{aligned}$$

So, $a_G \leq a_A$. Furthermore,

$$\begin{aligned} \log \frac{1}{a_H} &= \log \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{a_i} \right) = E \log \frac{1}{X} \geq E \left(\log \frac{1}{X} \right) = -\log(EX) \\ &= -\log a_G = \log \left(\frac{1}{a_G} \right). \end{aligned}$$

So, $a_G \geq a_H$.

定理 1.14 (Markov's (Chebyshev's) Inequality)

- If g is strictly increasing and positive on $(0, \infty)$, $g(x) = g(-x)$.
- X is a r.v. such that $E[g(X)] < \infty$, then for each $a > 0$

$$P(|X| \geq a) \leq \frac{E[g(X)]}{g(a)}$$



Proof:

$$\begin{aligned}
 E[g(X)] &\geq E[g(X)I_{\{g(X) \geq g(a)\}}] \\
 &\geq g(a)E[g(X)I_{\{g(X) \geq g(a)\}}] \\
 &= g(a)E[I_{|X| \geq a}] \\
 &= g(a)P(|X| \geq a)
 \end{aligned}$$

Some special cases: Markov's Inequality

$$g(x) = |x| \Rightarrow P(|X| \geq a) \leq \frac{E|X|}{a}$$

$$g(x) = x^p \Rightarrow P(|X| \geq a) \leq \frac{E|g(X)|}{a^p}$$

$$g(x) = x^2 \Rightarrow P(|X - EX| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

$$g(x) = e^{t|x|} \Rightarrow P(|X| \geq a) \leq \frac{E[e^{t|X|}]}{e^{ta}}$$

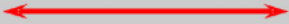
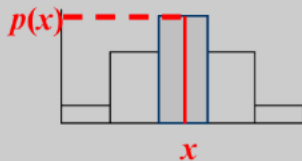

for some constant $t \geq 0$

第 1 讲 练习

1. If $\mu = EX \geq 0$ and $0 \leq \lambda \leq 1$, then

$$P(X > \lambda\mu) \geq \frac{(1 - \lambda)^2 \mu^2}{EX^2}$$

Consequently, if $E|Y| = 1$, $P(|Y| > \lambda) \geq (1 - \lambda)^2 / EY^2$ (This gives a lower bound complementing Chebyshev's inequality.)

RANDOM VARIABLE, X		
Type	Discrete	Continuous
Values	A finite/countable set of numbers x_1, x_2, x_3, \dots	All numbers in an interval 
Probability	Probability Mass Function, p <i>pmf</i> $P(X = x) = p(x)$ 	Probability Density Function, f <i>pdf</i> $P(a < X < b) = \left[\begin{array}{l} \text{area} \\ \text{under the} \\ \text{graph of } f \\ \text{over } (a, b) \end{array} \right]$ 

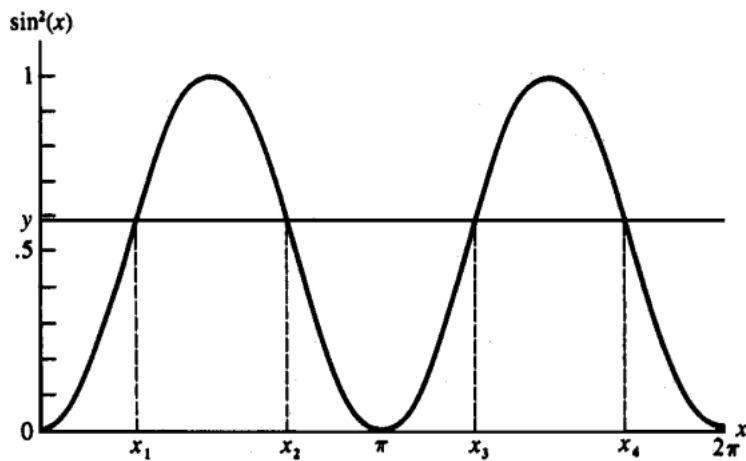
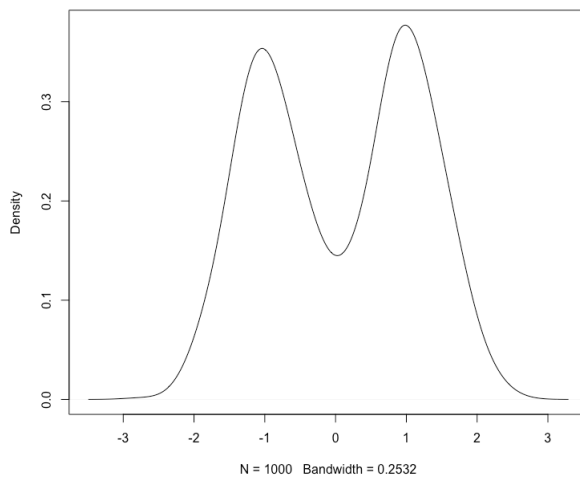
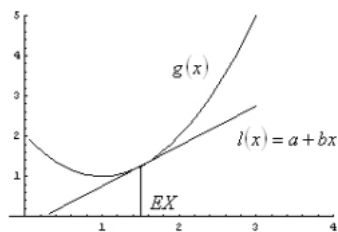


Figure 2.1.1. Graph of the transformation $y = \sin^2(x)$ of Example 2.1.2





第 2 讲 Convergence of Random Variables

内容提要

- ☐ converge in probability
- ☐ converge in L_p
- ☐ converge in quadratic mean
- ☐ almost sure converge
- ☐ converge in distribution;
- ☐ O_p, o_p

2.1 Look into Sample mean

- Recall:

Sample mean: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ Note: When $n \neq m$, X_n and X_m share the same expectation μ but have different distributions.

- Intuitively, when $n \rightarrow \infty$, \bar{X}_n is very close to $\mu = E(X)$.

```
2  n.vec <- seq(1, 10^3, 1)
3  n.len <- length(n.vec)
4  mean.full <- NULL
5  for(i in 1:n.len){
6    mean.full[i] <- mean(rnorm(n.vec[i]))
7  }
8  plot(n.vec, mean.full, type="l", xlab = "n",
9       ylab = "sample mean")
10 abline(h=0,lwd=1,col="blue")
```

If $x_1, x_2, \dots, x_n, \dots$ is an array of numbers, we know how to describe whether they converge or not. But what if they are random variables? How to describe it?

2.1.1 Generalization

- Let $\{Y_i\}_{i=1}^{\infty} = Y_1, Y_2, \dots, Y_n, \dots$ denotes a sequence of random variables
- Problem: How to describe the [limit](#) of Y_n
- Consider 2 cases:
 - Case 1. $Y_i \sim F$ independently, $i = 1, 2, \dots$

- Case 2. $Z_1 = Z_2 = Z_3 = \dots$, where $Z_1 \sim F$. Let $X \sim F$. Can we say $Y_i \rightarrow X$? Can we say $Z_i \rightarrow X$? How to differentiate these two cases?
- Recall: $Y_1, Y_2, \dots, Y_n : \omega \rightarrow R$. A sequence of functions

2.1.2 Convergence in Probability

定义 2.1 (Definition 5.5.1: Convergence in Probability)

For a sequence of r.v.'s $\{X_n\}_{i=1}^{\infty} = X_1, X_2, \dots, X_n, \dots$, we say they **converge in probability towards the r.v. X** (i.e. $X_n \xrightarrow{P} X$) if for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \epsilon) = 0$$

- The target X has **the same sample space** with all the X_i 's.
- X_n are usually dependent, but not identically distribution.
- Practically, find the sequence of events $A_n = \{\omega \in \Omega, |X_n(\omega) - X(\omega)| \geq \epsilon\}$ by obtaining **$|X_n - X|$ as a new r.v.**, and check if $P(A_n) \rightarrow 0$ when $n \rightarrow \infty$.

- Interpretation: for any ϵ , the event that $|X_n - X|$ has probability smaller than δ when n is large enough. It concerns more about the probability measure and r.v., instead of the CDF only.

定义 2.2 (random p-vectors)

For random p-vectors, X_1, X_n, \dots and X , if

$$\|X_n - X\| \xrightarrow{P} 0,$$

we say $X_n \xrightarrow{P} X$, where $\|z\| = (\sum_{i=1}^p z_i^2)^{1/2}$ denotes the Euclidean distance (L2-norm) for $z \in R^p$.



- It is easily to seen that $X_n \xrightarrow{P} X$ iff the corresponding component-wise convergence holds.

例题 2.1 Convergence in Probability

- Let X be a r.v. with prob 1 at 1, and $X_n \sim N(1, \frac{1}{n^2})$.

According to the property of normal distribution, $X_n - X \sim N(0, \frac{1}{n^2})$, so

$$\begin{aligned} P(|X_n - X| \geq \epsilon) &= P\left(|N(0, \frac{1}{n^2})| \geq \epsilon\right) \\ &\leq \frac{1}{n^2 \epsilon^2} \leq \delta, n \geq \frac{1}{\epsilon \sqrt{\delta}} \end{aligned}$$

So, $X_n \xrightarrow{P} X$.¹

例題 2.2 Convergence in Probability

- Let $X_n \sim \text{Ber}(0.5)$, and $X \sim \text{Ber}(0.5)$, X_n and X are independent. Note for any n ,

$$\begin{aligned} P(|X_n - X| \geq 1) &= P(\{X_n = 1, X = 0\} \cup \{X_n = 0, X = 1\}) \\ &= P(\{X_n = 1, X = 0\}) + P(\{X_n = 0, X = 1\}) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2} \not\rightarrow 0 \end{aligned}$$

So, X_n does NOT converge to X in probability.

¹Chebychev's inequality.

$$P(|X - \mu| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}$$

2.2 Convergence in L_r

定义 2.3 (Convergence in L_r)

For a sequence of r.v.'s $\{X_i\}_{i=1}^{\infty} = X_1, \dots, X_n, \dots$, we say they **converge in L_r towards the r.v. X** (i.e. $X_n \xrightarrow{L^r} X$) if for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} E|X_n - X|^r = 0.$$

where $[E(|X_n - X|^r)]^{\frac{1}{r}}$ is the L^r distance between X_n and X



- The target X has the same sample space with all the X_i 's
- When $r = 2$, converge in L^2 is also called converge in quadratic mean, i.e., $X_n \xrightarrow{qm} X$. The convergence in quadratic mean is generally used.
- To show L^r convergence, just figure out an upper bound of $E(|X_n - X|^r)$, and show this upper bound goes to 0.

例题 2.3 Example: Convergence in L_2

- Recall the previous example when X has a point mass at 1, and $X_n \sim N(1, \frac{1}{n^2})$. According to

the properties of normal distribution., $X_n - X \sim N(0, \frac{1}{n^2})$, so

$$\begin{aligned} E(|X_n - X|^2) &= (E(X_n - X))^2 + \text{Var}(X_n - X) \\ &= 0 + \frac{1}{n^2} = \frac{1}{n^2} \rightarrow 0. \end{aligned}$$

Hence, $X_n \xrightarrow{L^2} X$

- According to the deviation, if $\text{Var}(X_n - X) \rightarrow 0$, and $E(X_n - X) \rightarrow 0$, then there is

$$E(|X_n - X|^2) = (E(X_n - X))^2 + \text{Var}(X_n - X) \rightarrow 0$$

命题 2.1

If $\text{Var}(X_n - X) \rightarrow 0$, and $E(X_n - X) \rightarrow 0$, then $X_n \xrightarrow{L^2} X$.

命题 2.2

Let $0 < s < r < \infty$ if $X_n \xrightarrow{L^r} X$, then $X_n \xrightarrow{L^s} X$

- Recall that with [Holder inequality](#), there is

$$|E(XY)| \leq E|XY| \leq (E|X|^p)^{\frac{1}{p}} (E|Y|^q)^{\frac{1}{q}}$$

where $\frac{1}{p} + \frac{1}{q} = 1$

- Let $Y = 1$, $Z = |X_n - X|^r$, $l = r/s$, and $k = 1/(1 - s/r) > 1$, then

$$\begin{aligned} E(|X_n - X|^s) &= E(|X_n - X|^s \times 1) \\ &\leq \left[E(|X_n - X|^r) \right]^{s/r} \times 1^{1/k} \\ &= \left[E(|X_n - X|^r) \right]^{s/r} \rightarrow 0 \end{aligned}$$

命题 2.3

Let $0 < r < \infty$ if $X_n \xrightarrow{L^r} X$, then $X_n \xrightarrow{P} X$

Proof:

$$\begin{aligned} P(|X_n - X| \geq \varepsilon) &= P(|X_n - X|^r \geq \varepsilon^r) \\ &\leq \frac{E(|X_n - X|^r)}{\varepsilon^r} \rightarrow 0 \end{aligned}$$

Markov's Inequality : non-negative r.v.

$$P(X \geq a) \leq \frac{E(X)}{a}$$

定理 2.1 (Markov's (Chebyshev's) Inequality)

- If g is strictly increasing and positive on $(0, \infty)$, $g(x) = g(x)$.
- X is a r.v. such that $E[g(X)] < \infty$, then for each $a > 0$

$$P(|X| \geq a) \leq \frac{E[g(X)]}{g(a)}$$



Proof:

$$\begin{aligned} E[g(X)] &\geq E[g(X)I_{\{g(X) \geq g(a)\}}] \\ &\geq g(a)E[I_{\{g(X) \geq g(a)\}}] \\ &= g(a)E[I_{\{|X| \geq a\}}] \\ &= g(a)P(|X| \geq a) \end{aligned}$$

Some special cases: Markov's Inequality:

$$g(x) = |x| \rightarrow P(|X| \geq a) \leq \frac{E|X|}{a}$$

$$g(x) = x^p \rightarrow P(|X| \geq a) \leq \frac{E[g(X^p)]}{a^p}$$

$$g(x) = x^2 \rightarrow P(|X - EX| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

$$g(x) = e^{t|x|} \rightarrow P(|X| \geq a) \leq \frac{E[e^{t|X|}]}{e^{ta}}$$

for some constant $t \geq 0$

定义 2.4 (Definition 5.5.6)

For a sequence of r.v.'s $\{X_{i=1}^\infty\} = X_1, \dots, X_n, \dots$, we say they **almost sure convergence to r.v. X** (i.e. $X_n \xrightarrow{a.s} X$) if any $\epsilon > 0$,

$$P(\lim_{n \rightarrow \infty} X_n = X) = 1 \quad \text{or} \quad P(\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)) = 1$$



- The target X has **the same sample space** with all the X_i 's.
- $\{X_n\}$ and X are usually dependent

- Practically, to show the a.s. convergence,
 - For each outcome ω , find the sequence $X_1(\omega), X_2(\omega), \dots$ (sequence of real numbers) and the real number $X(\omega)$. Figure out whether $\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)$ is true or not.
 - Let the event $A = \{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}$.
 - Check if $P(A) = 1$
- Interpretation: for almost all the outcomes ω ! when n is large enough, $|X_n(\omega) - X(\omega)| \leq \epsilon$ for any $\epsilon > 0$

例题 2.4 Example 1: Almost Sure Convergence

- Let the sample space $\Omega = [0, 1]$, with a probability measure that is uniform on this space, i.e. $P([a, b]) = b - a$ for any $0 \leq a \leq b \leq 1$.

- Let

$$X_n(\omega) = \begin{cases} 1, & 0 \leq \omega < \frac{n+1}{2n} \\ 0, & \text{otherwise} \end{cases} \quad X(\omega) = \begin{cases} 1, & 0 \leq \omega < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

For each $\omega \in [0, 1]$.

- If $\omega \in [0, \frac{1}{2})$, then $X_n(\omega) = 1 = X(\omega)$.
- If $\omega = \frac{1}{2}$, then $X_n(\omega) = 1 \not\rightarrow X(\omega) = 0$.

- If $\omega \in (1/2, 1]$, then $X_n(\omega) = 0 = X(\omega)$, when $\frac{n+1}{2n} < \omega$, which is equivalent with $n \geq \frac{1}{2\omega-1}$.
So, $A = [0, 1/2) \cup (1/2, 1]$. Check $P(A) = 1$?

例题 2.5 Example 5.5.7: Almost Sure Convergence

- Let the sample space $\Omega = [0, 1]$, with a probability measure that is uniform on this space, i.e. $P([a, b]) = b - a$ for any $0 \leq a \leq b \leq 1$.
- Define r.v.

$$X_n(\omega) = \omega + \omega^n \quad \text{and} \quad X(\omega) = \omega$$

For each $\omega \in [0, 1]$.

- If $\omega \in [0, 1)$, $\omega^n \rightarrow 0$, then $X_n(\omega) \rightarrow \omega = X(\omega)$.
- If $\omega = 1$, then $X_n(\omega) = 2 \not\rightarrow X(\omega) = 1$ for every n
So, $A = [0, 1)$. Check $P(A) = 1$?
- Comparison between almost sure convergence and converge in probability
 - Convergence in probability: for each n , consider $P(|X_n(\omega) - X(\omega)| > \epsilon)$, and check the limit of this probability
 - Almost sure convergence: for each ω , check the limit $\lim_{n \rightarrow \infty} X_n(\omega)$, and find the probability of the set that the limit does not equal to $X(\omega)$

- Can we express it as the limit of probability?

定理 2.2 (Almost Sure Convergence)

The following statements are equivalent:

- $X_n \xrightarrow{a.s.} X$
- $\forall \epsilon > 0, P(\cap_{k \geq n} \{|X_k - X| < \epsilon\}) \rightarrow 1$
- $\forall \epsilon > 0, P(\cup_{k \geq n} \{|X_k - X| \geq \epsilon\}) \rightarrow 0$
- $\forall \epsilon > 0,$

$$\lim_{n \rightarrow \infty} P(\sup_{k \geq n} |X_k - X| > \epsilon) = 0$$



Here, we consider that set $\cup_{k \geq n} \{|X_k - X| > \epsilon\}$

命題 2.4

If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$



Proof: for any $\varepsilon > 0$

$$\begin{aligned} 0 &\leq P(|X_n - X| \geq \varepsilon) \\ &\leq P\left(\bigcup_{k=n}^{\infty} |X_k - X| \geq \varepsilon\right) \\ &= 0 \end{aligned}$$

Hence, $\lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) = 0$, which implies $X_n \xrightarrow{P} X$.

定义 2.5 (Definition 5.5.9)

Let $\{X_i\}_{i=1}^{\infty} = X_1, X_2, \dots, X_n, \dots$ be a sequence of r.v.'s with **CDF** F_1, \dots, F_n, \dots , and X be r.v. with **CDF** F . we say they **converges in distribution to r.v. X** (i.e. $X_n \xrightarrow{d} X$) if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

at **very point at which F is continuous.**



- $\{X_n\}$ and X can be dependent or independent
- Convergence:

- If X is discrete, the convergence stands at points F does not jump
- If X is cont., the convergence stands at every point
- Convergence in distribution is really the CDFs that converge, not the r.v. Hence it is quite different from conv. in prob. or alm. sure conv.

命题 2.5

If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$

Proof: Denote $F_n(x) = P(X_n \leq x)$ and $F(x) = P(X \leq x)$. First we have

$$\begin{aligned}
 F_n(x) &= P(X_n \leq x) \\
 &= P(X_n \leq x, |X_n - X| \leq \epsilon) + P(X_n \leq x, |X_n - X| > \epsilon) \\
 &\leq P(X \leq x - (X_n - X), |X_n - X| \leq \epsilon) + P(|X_n - X| > \epsilon) \\
 &\leq P(X \leq x + \epsilon) + P(|X_n - X| > \epsilon) \\
 &= F(x + \epsilon) + P(|X_n - X| > \epsilon)
 \end{aligned}$$

Or

$$\begin{aligned}
 F_n(x) &= P(X_n \leq x) \\
 &= P(X_n \leq x, X \leq x + \epsilon) + P(X_n \leq x, X > x + \epsilon) \\
 &\leq P(X_n \leq x, X \leq x + \epsilon) + P(|X_n - X| > \epsilon) \\
 &\leq P(X \leq x + \epsilon) + P(|X_n - X| > \epsilon) \\
 &= F(x + \epsilon) + P(|X_n - X| > \epsilon)
 \end{aligned}$$

On the other hand,

$$\begin{aligned}
 F_n(x) &= 1 - P(X_n \geq x) \\
 &= 1 - P(X_n \geq x, |X_n - X| \leq \epsilon) - P(X_n \geq x, |X_n - X| > \epsilon) \\
 &\geq 1 - P(X \geq x - (X_n - X), |X_n - X| \leq \epsilon) - P(|X_n - X| > \epsilon) \\
 &\geq 1 - P(X \leq x - \epsilon) - P(|X_n - X| > \epsilon) \\
 &= F(x - \epsilon) - P(|X_n - X| > \epsilon)
 \end{aligned}$$

Or

$$\begin{aligned}
 F_n(x) &= 1 - P(X_n \geq x) \\
 &= 1 - P(X_n > x, X > x - \epsilon) - P(X_n > x, X \leq x - \epsilon) \\
 &\geq 1 - P(X \geq x - \epsilon) - P(|X_n - X| > \epsilon) \\
 &\geq 1 - P(X \leq x - \epsilon) - P(|X_n - X| > \epsilon) \\
 &= F(x - \epsilon) - P(|X_n - X| > \epsilon)
 \end{aligned}$$

Combining the two, we have

$$F(x - \epsilon) - P(|X_n - X| > \epsilon) \leq F_n(x) \leq F(x + \epsilon) + P(|X_n - X| > \epsilon)$$

Letting $n \rightarrow \infty$ and since $X_n \xrightarrow{P} X$,

$$F(x - \epsilon) \leq \liminf_{n \rightarrow \infty} F_n(x) \leq \limsup_{n \rightarrow \infty} F_n(x) \leq F(x + \epsilon)$$

Recall that F is continuous at x , which means $F(x - \epsilon) \rightarrow F(x)$ and $F(x + \epsilon) \rightarrow F(x)$ as $\epsilon \rightarrow 0$. Hence,

$$F(x) \leq \liminf_{n \rightarrow \infty} F_n(x) \leq \limsup_{n \rightarrow \infty} F_n(x) \leq F(x)$$

Recall the characteristic function for $X \sim F$ is $\Phi_X(t) = E(e^{it})$. If $\Phi_X(t) = \Phi_Y(t)$ then X and Y have the same distribution.

定理 2.3

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of r.v.'s with characteristic functions $\Phi_{X_n}(t)$ and X be a r.v. with the characteristic function $\Phi_X(t)$. Then,

$$X_n \xrightarrow{d} X \Leftrightarrow \lim_{n \rightarrow \infty} \Phi_{X_n}(t) = \Phi_X(t)$$



Example: Suppose that $X_n \sim N(\mu + 1/n, \sigma^2 + 1/n)$, then

$$\Phi_{X_n}(t) = \exp\{(\mu + 1/n^2)t - t^2(\sigma^2 + 1/n)/2\} \rightarrow \exp\{\mu t - t^2\sigma^2/2\}$$

Note that the limit is the characteristic function for $X \sim N(\mu, \sigma^2)$.

So, $X_n \xrightarrow{d} X$. It is easier than the analysis on the CDF of X_n .

2.3 Relationship

定理 2.4 (Some comments)

•

$$X_n \xrightarrow{a.s.} X$$

$$\Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$$

$$X_n \xrightarrow{L_r}$$

- If $0 < s < r < \infty$, $X_n \xrightarrow{L_r} X \Rightarrow X_n \xrightarrow{L_s} X$.
- No other implications hold in general.



- (1).(a) If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$. **The converse may not hold.** Let

$$P(X_n = 0) = 1 - \frac{1}{n}, P(X_n = 1) = \frac{1}{n}$$

and X'_n 's are independent. Since $P(|X_n - 0| > \epsilon) = P(X_n = 1) = \frac{1}{n} \rightarrow 0$, Then $X_n \xrightarrow{P} 0$

X. However, $X_n \xrightarrow{a.s.} 0$ since for any $0 < \epsilon < 1$, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\cap_{k \geq n} \{|X_k - 0| < \epsilon\}) &= \lim_{n \rightarrow \infty} P(\lim_{r \rightarrow \infty} \cap_{k \geq n}^r \{|X_k| < \epsilon\}) \\ &= \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} P(\cap_{k \geq n}^r \{|X_k| < \epsilon\}) = \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} \prod_{k=n}^r (1 - \frac{1}{k}) \\ &= \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} \frac{n-1}{n} \frac{n}{n+1} \cdots \frac{r-1}{r} = \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} \frac{n-1}{r} = 0 \neq 1 \end{aligned}$$

(b) If $X_n \xrightarrow{L_r} X$, then $X_n \xrightarrow{P} X$. The converse may not hold.

$$P(X_n = 0) = 1 - \frac{1}{n}, P(X_n = n) = \frac{1}{n}$$

Then $X_n \xrightarrow{P} 0$ since

$$P(|X_n - 0| > \epsilon) = P(X_n = n) = \frac{1}{n} \rightarrow 0$$

. But $EX_n = 1 \neq 0$.

If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$. The converse may not hold.

$$X \sim N(0, 1), X_n = -X \sim N(0, 1)$$

Then $X_n \xrightarrow{d} X$. but $X_n \not\xrightarrow{P} X$ since

$$P(|X_n - X| > \epsilon) = P(2|X| > \epsilon) \not\rightarrow 0$$

. (2) If $0 < s < r < \infty$, $X_n \xrightarrow{L_r} X \Rightarrow X_n \xrightarrow{L_s} X$. The converse may not hold.

$$P(X_n = 0) = 1 - \frac{1}{n^2}, P(X_n = n) = \frac{1}{n^2}$$

Then $X_n \xrightarrow{L_1} X$ since

$$E|X_n - 0| = \frac{1}{n^2} \cdot n = \frac{1}{n} \rightarrow 0$$

But $X_n \not\xrightarrow{L_2} X$ since

$$E|X_n - 0|^2 = \frac{1}{n^2} \cdot n^2 = 1 \neq 0$$

(3). We now show that "a.s. convergence" and "mean convergence" do not imply each other.

- Let $P(X_n = 0) = 1 - n^{-2}$ and $P(X_n = n^3) = n^{-2}$. Then $X_n \xrightarrow{a.s.} 0$, but $X_n \not\xrightarrow{L_1} 0$. Since

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\cup_{k \geq n} \{|X_k - 0| \geq \epsilon\}) &= \lim_{n \rightarrow \infty} P(\lim_{r \rightarrow \infty} \cup_{k \geq n}^r \{|X_k| \geq \epsilon\}) \\ &= \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} P(\cup_{k \geq n}^r \{|X_k| \geq \epsilon\}) = \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} \sum_{k=n}^r \frac{1}{k^2} \\ &= \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} \left(\frac{1}{n^2} + \frac{1}{(n+1)^2} + \cdots + \frac{1}{r^2} \right) \rightarrow 0. \end{aligned}$$

However,

$$E|X_n - 0| = \frac{1}{n^2} \cdot n^3 \rightarrow \infty$$

- $X_n \xrightarrow{L_1} 0$, but $X_n \not\xrightarrow{a.s.} 0$

$$P(X_n = 0) = 1 - \frac{1}{n}, P(X_n = 1) = \frac{1}{n}$$

2.4 Properties of Convergence

- $X_n \rightarrow X$ and $Y_n \rightarrow Y$, then $X_n \pm Y_n \rightarrow X \pm Y$
 - $X_n \xrightarrow{a.s.} X, Y_n \xrightarrow{a.s.} Y$, then $X_n \pm Y_n \rightarrow X \pm Y$,

- $X_n \xrightarrow{L_r} X, Y_n \xrightarrow{L_r} Y$, then $X_n + Y_n \xrightarrow{a.s.} X + Y$,
- $X_n \xrightarrow{P} X, Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{L_r} X + Y$,
- $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} Y$, it is **not sure** that $X_n + Y_n \xrightarrow{d} X + Y$
- **Slutsky's Theorem** Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} C$, then
 - $X_n + Y_n \xrightarrow{d} X + C$
 - $X_n Y_n \xrightarrow{d} CX$
 - $X_n / Y_n \xrightarrow{d} X / C$ if $C \neq 0$
- **The Continuous Mapping Theorem.** If $g(\cdot)$ is a continuous function, then
 - $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$,
 - $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$,
 - $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

定理 2.5 (Continuous Mapping Theorem)

- Let X_1, X_2, \dots , and X be random p -vectors defined on a probability space
 - let $g(\cdot)$ be a vector-valued (including real-valued) continuous function defined on R^p .
- If X_n converges to X in probability, almost surely, or in law, then $g(X_n)$ converges to $g(X)$ in

probability, almost surely, or in law, respectively.



Remark The condition that $g(\cdot)$ is a continuous function in Theorem can be further relaxed to that $g(\cdot)$ is continuous a.s., i.e., $P(X \in C(g)) = 1$ where $C(g) = \{x : g \text{ is continuous at } x\}$ is called the continuity set of g .

例题 2.6

- If $X_n \xrightarrow{d} X \sim N(0, 1)$, then $1/X_n \xrightarrow{d} 1/X$?
- If $X_n = 1/n$, and

$$g(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Then $X_n \xrightarrow{d} g(X_n) \xrightarrow{d} ?$

2.5 Stochastic Orders

Recall:

- In mathematics, we use o and O notations to denote the order of terms
- $a_n = o(1)$ means $a_n \rightarrow 0$ when $n \rightarrow \infty$; $a_n = o(b_n)$ means that $a_n/b_n = o(1)$.
- $a_n = O(1)$ means $|a_n| \leq C$ for some constant $C > 0$, for all large n ; $a_n = O(b_n)$ means $a_n/b_n = O(1)$.

Now we consider the probabilistic version:

定义 2.6 (o_p)

If $X_n \xrightarrow{P} 0$, i.e. $P(|X_n| \geq \epsilon) \rightarrow 0$ for every $\epsilon > 0$, then we say that $X_n = o_p(1)$



定义 2.7 (O_p)

We say that $X_n = O_p(1)$, or X_n is bounded in probability, if for any $\epsilon > 0$, there exists $C_\epsilon > 0$, such that

$$P(|X_n| > C_\epsilon) \leq \epsilon$$

.



Generalisation: Consider a sequence X_1, X_2, \dots of r.v.'s and a_1, a_2, \dots , a sequence of positive

real numbers,

- For a r.v. $X, X_n \xrightarrow{P} X$ if only if $X_n - X = o_p(1)$
- $X_n = o_p(a_n)$ if only if $a_n^{-1} X_n = o_p(1)$. a_n is the rate.
- $X_n = O_p(a_n)$ if only if $a_n^{-1} X_n = O_p(1)$. a_n is the rate.

Examples:

- If $X_n \sim N(0, \frac{1}{n})$, then $X_n = o_p(1)$ and $X_n = O_p(\frac{1}{\sqrt{n}})$
- If $X_n = o_p(1)$, then $X_n = O_p(1)$
- $O_p(1)o_p(1) = o_p(1), O_p(1)O_p(1) = O_p(1)$
- $O_p(1) + o_p(1) = O_p(1)$
- $O_p(a_n)o_p(b_n) = o_p(a_nb_n), O_p(a_n)O_p(b_n) = O_p(a_nb_n)$
- $(1 + o_p)^{-1} = O_p(1)$
- $o_p(O_p(1)) = o_p(1)$

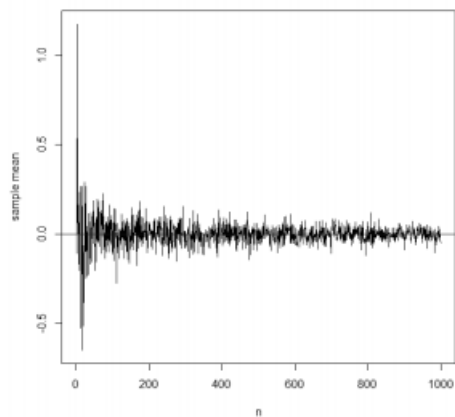
第2讲 练习

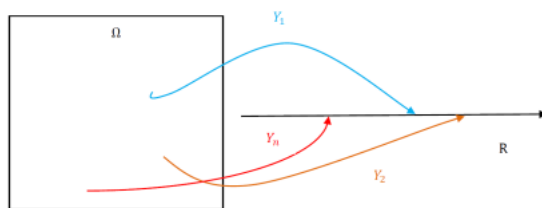
1. Show that if $X_n \xrightarrow{d} X$ for a random variable X , then $X_n = O_p(1)$.

2. Let X, X_1, X_2, \dots be a sequence of random variables. Show that $X_n \xrightarrow{P} X$ as $n \rightarrow \infty$ if and only if

$$E\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

3. Prove that $O_{P(1)} + o_{P(1)} = O_{P(1)}$.
4. Let X_1, X_2, \dots be iid random variables with $EX_1 = 0, EX_1^2 < \infty$, then $\sqrt{n}\bar{X}_n/S_n \xrightarrow{d} N(0, 1)$





第3讲 Law of Large Numbers and Central Limit Theorem

内容提要

- The Weak/Strong Law of Large Numbers
- The Central Limit Theorem

定理 3.1

Let X_1, X_2, \dots and X be random p-vectors,

1. **(The Portmanteau Theorem)** $X_n \xrightarrow{d} X$ is equivalent to the following condition:
 $E[g(X_n)] \rightarrow E[g(X)]$ for every **bounded continuous** function g .
2. **(Levy-Cramer continuity theorem)** Let $\phi_{X_1}, \phi_{X_2}, \dots, \phi_X$ be the character functions of X_1, X_2, \dots and X respectively. $X_n \xrightarrow{d} X$ iff $\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathcal{R}^p$

3. (Cramer-Wold device) $X_n \xrightarrow{d} X$ iff $c^\top X_n \xrightarrow{d} c^\top X$ for every $c \in \mathcal{R}^p$



例题 3.1

1. $X_n \sim \text{Uniform}\{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1\}$, $X_n \xrightarrow{d} X \sim ?$
2. If $g(x) = x^{10}$, then $E[g(X_n)] \rightarrow$

Proof:

1. For $\forall t \in [\frac{i}{n}, \frac{i+1}{n})$ Then

$$|F_{X_n}(t) - F_X(t)| = \left| \frac{i}{n} - t \right| < \frac{1}{n}$$

- 2.

$$E[g(X_n)] = \sum_{i=1}^n \frac{i^{10}}{n^{11}} \rightarrow E[g(X)] = \int_0^1 x^{10} dx = \frac{1}{11}$$

3.1 Law of Large Numbers

- Now that we have methods to describe the limit of a sequence of random variables

- Recall the motivating problem: Sample mean \bar{X}_n converges to EX intuitively.
- Question: what is this convergence? Is it convergence in distribution? probability? a.s.? L_r ?

3.1.1 Weak Law of Large Numbers (WLLN)

定理 3.2 (Theorem 5.5.1 WLLN)

Let $\{X_n\} = X_1, X_2, \dots$ be a sequence of independently and identically distributed (i.i.d.) r.v.'s such that $E|X_1| < \infty$, Then

$$\bar{X}_n \xrightarrow{P} EX_1$$



- The condition $E|X_1| < \infty$ is to assure the existence of EX_1 .
- The theorem can be extended to many dependence structures, such as Markov chains.
- The theorem can be extended to cases that X_i 's are not identical, but share the same 1st and 2nd moments.
- According to properties for convergence in probability, for any cont. function $g(\cdot)$,

$$g(\bar{X}_n) = g\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \xrightarrow{P} g(\mu)$$

- Sketch of Proof of WLLN

Note that when the limit is a constant, convergence in probability is equivalent with convergence in distribution. To prove convergence in distribution, we only need to show

$$\Phi_{X_n}(t) \rightarrow \Phi_{EX_1} = e^{itEX_1}$$

We will use the following result without any proof. For a r.v. X with finite first moment, we have

$$\Phi_X(t) = 1 + itEX + o(t)$$

Proof:

$$\begin{aligned}\Phi_{\bar{X}_n} &= E[\exp(it\bar{X}_n)] = E\left[\exp\left(it\frac{1}{n}\sum_{i=1}^n X_i\right)\right] \\ &= \prod_{i=1}^n E[\exp(itX_i/n)] \\ &= (E[\exp(itX_1/n)])^n = \Phi_{X_1}^n(t/n)\end{aligned}$$

Let $n \rightarrow \infty$, then

$$\Phi_{\bar{X}_n}(t) = (1 + itEX_1/n + o(1/n))^n \rightarrow e^{itEX_1}$$

Therefore, the convergence is proved.

3.1.2 Strong Law of Large Numbers (WLLN)

定理 3.3 (SLLN)

Let $X_n = X_1, X_2, \dots$, be a sequence of independently and identically distributed (i.i.d.) r.v.'s such that $E|X_1| < \infty$, then

$$\bar{X}_n \xrightarrow{a.s.} EX_1$$

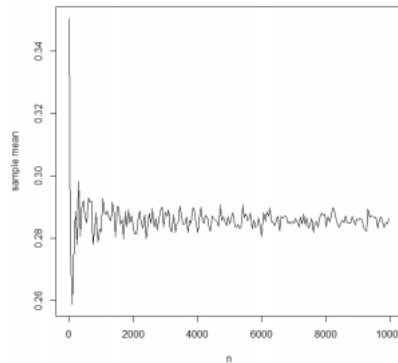


- The conditions can be relaxed. Identical distributions are not required, but there are still constraints on the second moment
- Stronger requirements than WLLN to assure better results The proof is beyond the scope of this course
- LLN: when n gets larger, the distribution of sample mean is more concentrated around EX_1 .

例题 3.2 Example of LLN: Calculate Expectation

Recall: $EX = \int x f(x) dx$, where $f(x)$ is pdf of X .

- Generate n samples with pdf $f(x)$, and calculate the \bar{X}_n . When n is very large, $EX \approx \bar{X}_n$
- Example: Beta distribution with parameters $a = 2$ and $b = 2$, $EX = \int_0^1 \frac{\tau(7)}{\tau(2)\tau(5)} x^{2-1} (1-x)^{5-1} dx$. **Hard to Calculate!**



```

1  rm(list=ls())
2  n.vec <- seq(1, 10^4, 50)
3  n.len <- length(n.vec)
4  mean.full <- NULL
5  for(i in 1:n.len){
6  mean.full[i] <- mean(rbeta(n=n.vec[i], shape1=2,shape2=5))
7  }
8  plot(n.vec, mean.full, type="l",xlab = "n",
9  ylab = "sample mean")
10 abline(h=2/7,lwd=1,col="blue")

```

- What's more, according to the continuous mapping theorem, $g(\bar{X}_n) \rightarrow g(E(X))$, e.g., $[E(X)]^2 \approx \bar{X}_n^2$

例题 3.3 Example of LLN: Calculate Expectation

- LLN can also be used to find $E(g(X))$, where $g(\cdot)$ is a function
- Generate n i.i.d. samples $\{X_i\}_{i=1}^n$ with pdf $f(x)$, and let $Y_i = g(X_i)$. Then $\bar{Y}_n \rightarrow E[g(X)]$. When n is very large, $E(g(X)) \approx \bar{Y}_n$
- Example: Beta distribution with parameters $a = 2, b = 5$. $Y = X^2, Z = 2X + 1, W = e^X$

例题 3.4 Examples of Using LLN: Integration

- Suppose we wish to calculate

$$\int_0^1 g(x) dx$$

where $g(x)$ may be complicated and the integration is not easy to compute.

- Relate the integration with expectation. We need a density function. Let $X \sim \text{Unif}(0, 1)$, then the pdf of x is 1 on $[0, 1]$. For function $g(\cdot)$,

$$Eg(X) = \int_0^1 g(X) \cdot 1 dx = \int_0^1 g(x) dx$$

Procedure (apply the method in previous slide for mean):

- Generate n i.i.d samples $X \sim \text{Unif}(0, 1)$, and calculate $g(X_i)$ correspondingly
- Compute $Eg(X) \sim g(\bar{X}_i) = \frac{1}{n} \sum_{i=1}^n g(X_i)$
- This method is called **Monte Carlo** method.

3.2 Central Limit Theorem (CLT)

3.2.1 Motivation

Suppose that a fair coin is tossed 100 times. What is the probability that the total number of heads is no smaller than 60?

Let X be the total number of heads, then $X \sim \text{Bin}(100, 0.5)$. We are interested in $P(X \geq 60)$

- Calculate directly means calculating 40 probs $\{p(X = i)\}_{i=60,61,\dots}$ and take the summation. **COMPLICATED.**
- X can be seen as the summation of 100 Bernoulli trials with $p = 0.5$ and limit theorems can be applied.
 - With LLN, we only know $X/100 \xrightarrow{P} 0.5$, CANNOT get $P(X \geq 60)$
 - New Limit Theorem is required to **describe the behaviour of X more accurately.**

定理 3.4 (Theorem 5.5.14 Central Limit Theorem (CLT))

Let $\{X_n\} = X_1, X_2, \dots$ be a sequence of independently and identically distributed (i.i.d.) r.v.'s

s such that $EX_1^2 < \infty$. Let $\sigma^2 = \text{Var}(X_1)$ and $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, then

$$\sqrt{n}[\bar{X}_n - EX_1] = \sqrt{n}\left[\frac{1}{n} \sum_{i=1}^n (X_i - EX_1)\right] \xrightarrow{d} N(0, \sigma^2)$$

- $EX_1^2 < \infty$ is a regular condition to assure the existence of EX_1 and $\text{hVar}(X_1)$
- It means that \bar{X}_n can be approximated by a normal distribution, no matter what the distribution for X_i is.
- Here, $n^{-0.5}$ is the convergence rate. Or, say, $\bar{X}_n - EX_1 = O_p(n^{-0.5})$. If we use $n^{0.5+\delta}$ with $\delta > 0$, then there is no meaningful result; if we use $n^{0.5-\delta}$ then it converges to 0.



注 CLT is the most important theorem in statistics

- CLT means that, the sample mean will be approximately normally distributed for large sample sizes, regardless of the distribution of the samples
- Many statistics (say, \bar{X}_n , \bar{X}_n^2) have distributions that are approximately normal, even the population distribution is not normal (\Leftarrow The dist. of statistics can be approximated)
- Statistical inference can be derived based on normality, provided the sample size is large

- In practice, it gives a very rough guideline to approximate \bar{X}_n when n is large (a few hundreds or even more)
- However, the convergence is the weakest convergence, converge in distribution. With the result, for statistics (e.g., \bar{X}_n), we can only calculate

$$P(\bar{X}_n \geq a), P(\bar{X}_n \leq a), P(a \leq \bar{X}_n \leq b)$$

3.2.2 Comparison Between LLN and CLT

LLN		CLT
Results	Focus on \bar{X}_n	Focus on $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$
Convergence	In probability	In distribution
Interpretation	\bar{X}_n converges to μ	The rate \bar{X}_n converges to μ
Usage	Monte Carlo Method	Statistical Inference

表 3.1: The differences of LLN and CLT

- Compare to real numbers, LLN means that

$$\frac{2\sqrt{n} + 1}{\sqrt{n}} \rightarrow 2$$

CLT mean that

$$\sqrt{n}\left(\frac{2\sqrt{n}+1}{\sqrt{n}}-2\right) \rightarrow 1$$

例题 3.5 CLT Example

Let $X_i \sim i.i.d.Exp(1), i = 1, 2, \dots$. We know that $E[X_1] = Var(X_1) = 1$, and so the sample mean converges to 1. How many samples we need so that our error is at most 10

The target is, to figure out n , so that $P(0.9 \leq \bar{X}_n \leq 1.1) \geq 0.95$. For large n , with CLT, we have $\sqrt{n}(\bar{X}_n - 1) \xrightarrow{d} N(0, 1)$. Therefore, we may use standard normal distribution to approximate the probability. Then,

$$\begin{aligned} P(0.9 \leq \bar{X}_n \leq 1.1) &= P(-0.1 \leq \bar{X}_n - 1 \leq 0.1) \\ &= P(-0.1\sqrt{n} \leq \sqrt{n}(\bar{X}_n - 1) \leq 0.1\sqrt{n}) \\ &\approx \Phi(0.1\sqrt{n}) - \Phi(-0.1\sqrt{n}) \\ &= 2\Phi(0.1\sqrt{n}) - 1 \geq 0.95 \end{aligned}$$

Check the normal table, and we can find $n > 384$.

第3讲 练习

1. Suppose X_1, \dots, X_n are iid with mean μ , variance σ^2 , and $EX_1^4 < \infty$, then

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \mu_4 - \sigma^4)$$

where μ_4 denotes the centered fourth moment of X_1 .

2. (Multivariate CLT for iid case) Let \mathbf{X}_i be iid random p -vectors with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} N_p(0, \boldsymbol{\Sigma})$$

第 4 讲 Edgeworth Expansion and Models

内容提要

- ☐ Delta Method
- ☐ Edgeworth Expansion
- ☐ Multivariate Delta method

4.1 The Delta Method

- CLT is for \bar{X}_n
- Can we generalize it to $g(\bar{X})$? So the application of CLT is extended.
- Example: what is the limiting dist. for $e^{\bar{X}_n}$?

定理 4.1 (Theorem 5.5.24: The Delta Method)

Suppose that $\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$, and that $g(\cdot)$ is a **differentiable** function such that $g'(\mu) \neq 0$, then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{d} N(0, 1)$$

**注**

- $Y_n = \bar{X}_n$ for the CLT. Here, Y_n is a generalized case. For any Y_n satisfying the convergence rule, we have the delta method.
- In other words, $Y_n \xrightarrow{d} N(\mu, \sigma^2/n)$ implies that $g(Y_n) \xrightarrow{d} N(g(\mu), (g'(\mu))^2 \sigma^2/n)$

Intuition of proof: We are interested in the term $\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma}$. Since $\frac{\sqrt{n}(Y_n - \mu)}{\sigma}$ converges to normal distribution, we have that $Y_n - \mu = O_p(n^{-0.5})$. Since g is differentiable at μ , when $y - \mu = o(1)$, there is

$$g(y) = g(\mu) + g'(\mu)(y - \mu) + o(1)(y - \mu)$$

Introduce it into the term of interest,

$$\begin{aligned}
 \frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} &= \frac{\sqrt{n}(g(\mu) + g'(\mu)(Y_n - \mu) + rem - g(\mu))}{|g'(\mu)|\sigma} \\
 &= \frac{\sqrt{n}(g'(\mu)(Y_n - \mu) + rem)}{|g'(\mu)|\sigma} \\
 &= \sqrt{n} \left[\frac{Y_n - \mu}{\sigma} \right] + \frac{\sqrt{n}}{|g'(\mu)|\sigma} \cdot rem
 \end{aligned}$$

The first term converges to $N(0, 1)$ in distribution, and **the second term converges to 0**. So the summation would converge to $N(0, 1)$ in distribution.

Note: The point here is that $g(y) = g(\mu) + g'(\mu)(y - \mu) + o(1)(y - \mu)$ when y changes with a small quantity. Since **delta (δ)** is always used to denote small quantity, so the method is called **”The Delta Method”**.

定理 4.2

Suppose that $\mathbf{Y}_n = (Y_{n1}, Y_{n2}, \dots, Y_{nk})$ is a sequence of random vectors such that $\sqrt{n}(\mathbf{Y}_n - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$. Let $g : \mathcal{R}^k \rightarrow \mathcal{R}^m$ be once differentiable at $\boldsymbol{\mu}$ with the gradient matrix $\Delta_g(\boldsymbol{\mu})$,

then

$$\sqrt{n}(g(\mathbf{Y}_n) - g(\boldsymbol{\mu})) \xrightarrow{d} N_m(\mathbf{0}, \Delta_g^\top(\boldsymbol{\mu}) \boldsymbol{\Sigma} \Delta_g(\boldsymbol{\mu}))$$

provided $\Delta_g^\top(\boldsymbol{\mu}) \boldsymbol{\Sigma} \Delta_g(\boldsymbol{\mu})$ is positive definite.



例题 4.1 Example 1: The Delta Method Let X_1, \dots, X_n be i.i.d with finite mean μ and finite variance σ^2 . By the CLT,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

If $W_n = e^{\bar{X}_n}$, $W_n \xrightarrow{d} ?$

Proof: $W_n = g(\bar{X}_n)$ where $g(t) = e^t$. Since $g'(t) = e^t$, the delta method implies that

$$\frac{\sqrt{n}(W_n - e^\mu)}{\sigma e^\mu} \xrightarrow{d} N(0, 1)$$

例题 4.2 Example 2: The Delta Method Let X_1, \dots, X_n be i.i.d with finite mean μ and finite covari-

ance matrix Σ . The according to the multivariate CLT, we have

$$\sqrt{n} \left(\begin{bmatrix} \bar{X}_{n1} \\ \bar{X}_{n2} \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right) \xrightarrow{d} N_2(\mathbf{0}, \Sigma)$$

Let $g(s, t) = s^2 + t^2$, then can apply the delta method and we have that

$$\sqrt{n} \left(\bar{X}_{n1}^2 + \bar{X}_{n2}^2 - \mu_1^2 - \mu_2^2 \right) \xrightarrow{d} 2\mu_1 Z_1 + 2\mu_2 Z_2$$

where $Z = (Z_1, Z_2)^T \sim N(0, \Sigma)$.

4.2 The Edgeworth Expansion

- Note that CLT is a good estimate for large n
- When n is small, it might be away from the truth
- Example: Consider $X_i \stackrel{i.i.d}{\sim} \text{Gamma}(1, 1)$, $i = 1, 2, 3$, We know that $EX_1 = \text{Var}X_1 = 1$. Let $Z = \sqrt{3}(\bar{X}_3 - 1)$, then according to CLT, $Z \sim N(0, 1)$ approximately.
- Since n is small, there is large difference! How to get a good estimate?

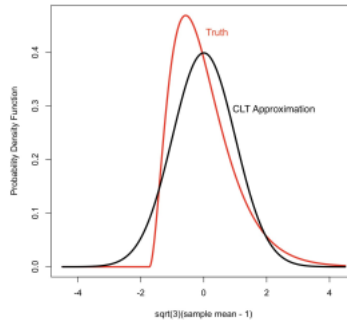


图 4.1

Recall that in the proof of CLT, we figured out the characteristic function as

$$\Phi_{Z_n}(t) = \Phi_Y^n\left(\frac{t}{\sqrt{n}}\right)$$

where $Z_n = \frac{1}{\sigma} \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n (X_i - EX_i) \right]$ and $Y = (X_1 - EX_1)/\sigma$. When we take it as $\Phi_Y(t/\sqrt{n}) = 1 - \sigma^2 t^2 / (2\sigma^2 n) + o(1/n)$, then we have CLT.

What if we consider higher-order moments?

- Let k_j be j -th cumulant of Y , where

$$k_1 = EY, k_2 = \text{Var}(Y), k_3 = E(Y - EY)^3, k_4 = E(Y - EY)^4 - 3\text{Var}Y$$

- Consider $K_Y = \log(\Phi_Y(t))$ and the Taylor series expansion of $K_Y(t)$ at $t = 0$,

$$K_Y(t) = \frac{1}{\Phi_Y(0)} \frac{d\Phi_Y(t)}{dt} \Big|_{t=0} + \frac{1}{\Phi_Y(0)} \frac{d^2\Phi_Y(t)}{dt^2} \Big|_{t=0} + \cdots = \sum_{j=1}^{\infty} k_j \frac{(it)^j}{j!}$$

- Since $Y = (X_1 - EX_1)/\sigma$, we have $k_1 = EY = 0, k_2 = \text{Var}Y = 1$, so

$$\Phi_Y(t) = \exp\left(-\frac{t^2}{2} + \sum_{j=3}^{\infty} \frac{(it)^j}{j!}\right) = \exp\left(-\frac{t^2}{2} + \frac{k_3^3(it)}{3!} + \cdots + \frac{k_j^j}{j!} + \cdots\right)$$

- Introduce it into the equation for

$$\Phi_{Z_n}(t) = e^{-t^2/2} [1 + n^{-1/2} r_1(it) + n^{-1} r_2(it) + o(n^{-1})]$$

where $r_1(it) = k_3^3(it)/6$, and $r_2(it) = \frac{1}{24} k_4^4(it) + \frac{1}{72} k_3^{12}(it)$

- Here, we apply the higher order moments of X to depict the characteristic function more clearly.

Therefore, we have the Edgeworth expansion:

$$F_{Z_n}(z) = \Phi(z) + \frac{1}{\sqrt{n}} p_1(z) \phi(z) + \frac{1}{n} p_2(z) \phi(z) + o\left(\frac{1}{n}\right)$$

where $p_1(z) = -k_3(z^2 - 1)/6$ and

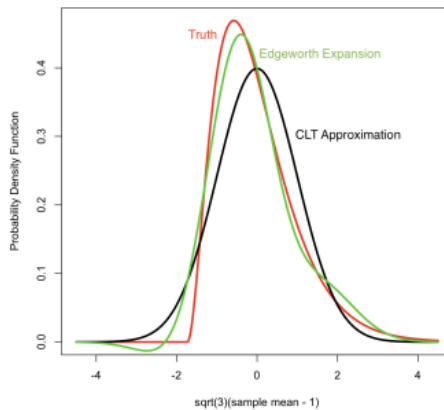
$$p_2(z) = -z[k_4(z^2 - 3)/24 + k_3^2(z^4 - 10z^3 + 15)/72]$$

.

Returning to our example. Consider the r st order Edgeworth expansion, we have

$$P(Z_e \leq z) \approx \Phi(z) + \frac{1}{\sqrt{3}\phi(z)p_1(z)}$$

where $p_1(z) = -k_3(z^2 - 1)/6$ and $k_3 = 2$.



第4讲 练习

1. Let X_1, \dots, X_n, \dots be a sequence of independent random variables such that

$$E[X_i] = \mu, \quad \text{Var}(X_i) < \sigma^2, \quad n = 1, 2, \dots$$

With Chebychev's inequality, prove that $\bar{X}_n \xrightarrow{P} \mu$.

2. Let U_1, U_2, \dots be independent random variables having the uniform distribution on $[0,1]$ and $Y_n = (\prod_{i=1}^n U_i)^{-1/n}$. Show that $\sqrt{n}(Y_n - e) \xrightarrow{d} N(0, e^2)$.
3. Let X_1, \dots, X_n be i.i.d. random variables following Uniform $[0,1]$. Let $Y_n = \min(X_1, \dots, X_n)$.
(i) Show that $Y_n \xrightarrow{a.s} 0$ as $n \rightarrow \infty$. (ii) Show that $nY_n \xrightarrow{d} \exp(1)$, where $\exp(1)$ is the exponential distribution with density $f(x) = e^{-x}$ for $x > 0$.

第5讲 Principle of Data Reduction

内容提要

- Parametric models
- Data reduction via statistics
- How to construct estimators

- WLLN and CLT shows that the sample average is a reasonable estimator for the expectation
 - Converges to the expectation
 - Rate $O(1/\sqrt{n})$
- Is sample average the best estimator for the expectation?
 - 'Best' in what sense?
 - If not, how to find the 'best' estimate?
 - What performance will the 'best' estimate have?
- Estimations for other parameters, or function of parameters?
 - Example: $X \sim N(\mu, \sigma^2)$. What is the estimation for σ ?

- Example: $X \sim \text{Gamma}(\alpha, \beta)$. How to estimate α and β ? How about $\alpha + \beta$?
- Not all of them can be estimated from the sample mean
- What is a proper estimation?

5.1 Introduction

This section covers the section topic of our class. including:

- Parametric models
- Data reduction via statistics
- How to construct estimators


Evaluation of estimators will be covered in the third topic.

5.2 Population and Sample

- Population

- The collection of measurements on a variable of interest: e.g., the condition of each light bulb of one manufactory.
- Usually, hypothesize a model: e.g, Bernoulli(p)
- Sample

定义 5.1 (Definition 5.1.1: Random sample)

The random variables X_1, X_2, \dots, X_n are called a random sample of size n from the population $f_X(x)$ if X_1, X_2, \dots, X_n are i.i.d. random variables with PMF or PDF $f_X(x)$. 

- Example: A sample from the light bulb manufactory: $X_1, X_2, \dots, X_n \sim \text{Ber}(p)$
- The "i.i.d" condition can be relaxed
- If "i.i.d" condition holds, then the joint density of the random sample is

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i)$$

5.3 Parameter Estimation

- Usually, $f_X(\cdot)$ is not known to us. We draw samples to explore the properties of $f_X(\cdot)$, e.g., expectation, variance, tails, etc.
- If prior information is known, say, f_X has an unknown **finite dimensional parameter** $\theta \in \Theta$, which characterizes f_X . Then the problem is to estimate θ

- Joint distribution of the sample:

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_X(x_i; \theta)$$

- Estimate θ
- Construct some statistical tests for θ ; say, model selection
- Asymptotic properties
- If no prior information is known, we cannot assume the distribution family for f_X . We call it as **non-parametric statistics**
 - Splines
 - Kernel estimation
 - etc.

例題 5.1 I.I.D. Normal Model

- A basic model statisticians usually use is the normal model.
- Let $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$. Here, the unknown parameters are $\theta = (\mu, \sigma^2)$.
- Given the observations x_1, \dots, x_n , the joint density is

$$f_{X_{1:n}}(x_1, \dots, x_n; \theta) = \prod_{i=1}^n \frac{\exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}}{\sqrt{2\pi}\sigma} = \frac{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}}{(\sqrt{2\pi}\sigma)^n}$$

- If the observations are given, then $f_{X_{1:n}}(x_1, \dots, x_n; \theta)$ can be seen as a function about θ ,

$$L(\theta; x_1, \dots, x_n) = \frac{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}}{(\sqrt{2\pi}\sigma)^n}$$

$L(\theta; x_1, \dots, x_n)$ is called the **likelihood function** for this models.

- One way to estimate the parameters $\theta = (\mu, \sigma^2)$ is to find the maximister of $L(\theta)$:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta; x_1, \dots, x_n)$$

Maximum likelihood function estimation for normal dist.(Quick review)

- Note that $l(\theta) = \log L(\theta)$ has the same maximizer with $L(\theta)$.

$$l(\theta; x_1, \dots, x_n) = \log L(\theta; x_1, \dots, x_n) = \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} - \frac{n}{2} \log(2\pi\sigma^2)$$

- Take the partial derivative, we have

$$\left(\frac{\partial l(\theta)}{\partial \mu}, \frac{\partial l(\theta)}{\partial \sigma^2} \right) = \left(\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu), -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu)^2 \right) =$$

- Let the derivative to be 0 (local extrema). The solution is

$$\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad \tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \tilde{\mu}_n)^2$$

- Since this is the only solution, this local extrema should be a global extremum. Check whether it is a maxima. We need the Hessian matrix to be a **negative definite matrix**.

例题 5.2 The Exponential Family

- Generalize the family of normal distribution
- **Exponential family** is a class of densities, for a random variable X and parameter θ , the density function is

$$f_X(x; \theta) = h(x) \exp\{\eta(\theta)T(x) - A(\theta)\}$$

- h, T, A are known functions
- The density functions is a product of data-only part ($h(x)$), parameter-only part $\exp\{-A(\theta)\}$, and the cross-term of data and parameters.
- The cross-term can be expressed as the exponential transformation of the product of parameter and data.
- Joint density:

$$f(x_1, \dots, x_n) = \left(\prod_{i=1}^n h(x_i) \right) \exp \left\{ \eta(\theta) \sum_{i=1}^n T(X_i) - nA(\theta) \right\}$$

[The Exponential Family-Example]

- The normal distribution belongs to the exponential Family. If $X \sim N(\mu, \sigma^2)$, the density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2 - \frac{\mu^2}{2\sigma^2} - \log \sigma}\right\}$$

Let $h(x) = \frac{1}{\sqrt{2\pi}}$, $\eta(\theta) = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$, $T(x) = (x^2, x)$, $A(\theta) = \frac{\mu^2}{2\sigma^2} + \log \sigma$. Then we have

$$f_X(x; \theta) = h(x) \exp\{\eta(\theta)^T T(x) - A(\theta)\},$$

which is an exponential family distribution.

5.4 Some popular models

5.4.1 Bayesian Models: Example

Example: Suppose that $X_i|\theta \sim \exp(\lambda)$. Also, we know that $\lambda \sim \text{Gamma}(a, b)$. Given the observations x_1, x_2, \dots, x_n , what information can we get about λ ?

Solution. The joint density for x_1, x_2, \dots, x_n and λ is

$$\begin{aligned} f(x_1, x_2, \dots, x_n, \lambda) &= \left[\prod_{i=1}^n f_X(x_i; \lambda) \right] \pi(\lambda) \\ &= \left[\prod_{i=1}^n \lambda e^{-\lambda x_i} \right] \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{\lambda \sum x_i} \lambda^{a-1} e^{-b\lambda} \\ &= \frac{b^a}{\Gamma(a)} \lambda^{n+a-1} e^{-\lambda(b + \sum x_i)} \end{aligned}$$

Now, we are curious about λ , so we want to know the conditional distribution of λ given the observations. According to the definition of conditional distribution, we have

$$\pi(\lambda|x_1, x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n, \lambda)}{f(x_1, x_2, \dots, x_n)}$$

Here, to differentiate the density function for X and λ , we always use π for the density function of λ , and f for the density function of X_1, \dots, X_n .

例题 5.3 Bayesian Models: Example

We want to solve

$$\pi(\lambda|x_1, x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n, \lambda)}{f(x_1, x_2, \dots, x_n)}$$

- The numerator is known, $f(x_1, x_2, \dots, x_n, \lambda) = \frac{b^a}{\Gamma(a)} \lambda^{n+a-1} e^{-\lambda(b+\sum x_i)}$
- The denominator can be calculated :

$$f(x_1, x_2, \dots, x_n) = \int_{\lambda} f(x_1, x_2, \dots, x_n, \lambda) d\lambda = \frac{b^a}{\Gamma(a)} \int_0^{\infty} \frac{b^a}{\Gamma(a)} \lambda^{n+a-1} e^{-\lambda(b+\sum x_i)} d\lambda = \frac{b^a}{\Gamma(a)} \times \frac{\Gamma(b+\sum x_i)}{(b+\sum x_i)^{b+\sum x_i}}$$

- So, the conditional distribution for λ is

$$\begin{aligned} \pi(\lambda|x_1, x_2, \dots, x_n) &= \frac{(b + \sum_{i=1}^n x_i)^{n+a}}{\Gamma(n+a)} \frac{b^a}{\Gamma(a)} \lambda^{n+a-1} e^{-\lambda(b+\sum x_i)} \\ &\sim \text{Gamma}(n+a, b + \sum x_i) \end{aligned}$$

We call it as **posterior distribution**.

- In the previous example, if we have prior information about λ , say, the expectation and variance, then we can identify the values of a and b , and the posterior is totally known.
- In Bayesian statistics, the posterior function is the "final answer". For frequentist, the estimation is a value.
- Depend on the loss function, the posterior function can be further reduced to a value. For example, when the loss function is $L^2 = \text{loss}((\hat{\theta} - \theta)^2)$, then the Bayes estimator can be reduced as $E[\pi(\theta|x_1, \dots, x_n)]$. Details discussed later.
- Therefore, there is no "confidence interval" in Bayesian statistics. A similar notion is "credible interval". Details later.

5.4.2 The Linear Model

- Consider a sequence of data in pairs: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where $y_i \in \mathbb{R}$, and $x_i \in \mathbb{R}^p$, $p \geq 1$, $1 \leq i \leq n$.
- The y_i 's are called **response variables** and the x_i 's are **explanatory variables**. It is hypothesised that there exist some functional relationship of the form

$$Y_i = g_{\theta}(x_i) + \epsilon$$

Therefore, we can use x_i to predict the responses y_i . Here, ϵ_i is interpreted as noise.

- A simple prediction function is linear function. Therefore, the prediction function is

$$Y_i = \theta_0 + \sum_{j=1}^{p-1} \theta_j x_{ij} + \epsilon_i,$$

where ϵ_i are i.i.d. zero mean random samples, usually assumed to be $N(0, \sigma^2)$. This model is called the [linear regression model](#).

5.5 Statistics

- Random sample: X_1, X_2, \dots, X_n .
- Work on the random sample to achieve information.

定义 5.2 (Definition 6.2.1 Statistics)

For a random sample X_1, X_2, \dots, X_n , a [statistics](#) is a function of the random sample $T(X_1, X_2, \dots, X_n)$.



- The statistic $T(X)$ is also a random variable. Most times, its distribution changes with n , and we denote the CDF as G_n , called the **sampling distribution**.
- With the observations x_1, x_2, \dots, x_n , we have $T(x_1, x_2, \dots, x_n)$, a **realization** of the statistic $T(X_1, X_2, \dots, X_n)$.

例题 5.4 Statistics: Examples

Some examples of Statistics:

- Single observation of the sample: X_1
- order statistics: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ (the bracket means we are considering the order)
- Sample mean: \bar{X}_n .
- Sample variance: $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
- sample minimum: $X_{(1)}$
- sample maximum: $X_{(n)}$
- sample range: $X_{(n)} - X_{(1)}$

what is a "good" statistics?

5.5.1 Properties of Statistics

- Recall that we are doing parametric inference, where the model is

$$f_X(x; \theta)$$

- We hope the statistics can be a summary of all the data, relevant to the parameter.
- The process can be seen as a **data reduction** process

Properties:

- Sufficient statistics
- Ancillary statistics
- Complete statistics

5.5.2 Sufficiency

- Reduce the data, so that all the information relevant to the parameter can be summarized in one statistic

Sufficiency principle

Let $X = (X_1, X_2, \dots, X_n)$ be random sample from the distribution $f(x; \theta)$. If $T(x)$ is a sufficient statistic for θ , then any inference about θ should depend upon the sample X **only through the value** of $T(X)$

- We say, $T(X)$ is sufficient for the parameter θ
- We can replace X with $T(X)$ without losing information

定义 5.3 (Definition 6.2.1: Sufficiency/sufficient statistics)

A statistic $T(X)$ is a sufficient for θ if the conditional distribution of the sample X given $T(X)$ does not depend on θ , i.e.,

$$f(x_1, x_2, \dots, x_n | t; \theta) = f(x_1, x_2, \dots, x_n | t).$$



The above definition is not easy to check whether a statistic $T(\mathbf{X})$ is a sufficient statistic.

Theorem 6.2.2

If $p(\mathbf{x}|\theta)$ is the pdf or pmf of \mathbf{X} , and $q(\mathbf{t}|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every \mathbf{x} ,

$$\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} \equiv \text{constant} \quad \text{in } \theta$$

例题 5.5 Example 1 $X_1, X_2, \dots, X_n \sim \text{Poisson}(\theta)$. Let $T = \sum_{i=1}^n X_i$. Since Poisson distribution is a discrete distribution, we are working with the PMF. The conditional distribution is

$$P(x_1, x_2, \dots, x_n | t) = \frac{P(X_1=x_1, \dots, X_n=x_n, T=t)}{P(T=t)}$$

Since $T = \sum_{i=1}^n X_i$,

$$P(X_1 = x_1, \dots, X_n = x_n, T = t) = \begin{cases} 0, & T(x) \neq t \\ P(X_1 = x_1, \dots, X_n = x_n), & T(x) = t \end{cases}$$

And,

$$P(X^n = x^n) = \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{\prod (x_i!)} = \frac{e^{-n\theta} \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n \prod (x_i!)}$$

Now, $T(x_1, \dots, x_n) = \sum x_i = t$. According to the property of Poisson distribution, $T \sim \text{Poisson}(n\theta)$,

$$P(T = t) = e^{-n\theta} (n\theta)^t / t!$$

so

$$P(X^n = x^n) / P(T = t) = t! / \left[\prod (x_i!) n^t \right]$$

which does not depend on θ . So, T is a sufficient statistic for θ .

例题 5.6 Example 2 Let X_1, X_2, \dots, X_n be i.i.d. *Bernoulli*(p). Let $T(\mathbf{X}) = X_1 + X_2 + \dots + X_n$.

Then

$$p(\mathbf{x}|p) = p^{x_1 + \cdots + x_n} (1 - p)^{n - (x_1 + \cdots + x_n)}$$

$$q(t|p) = \binom{n}{t} p^t (1 - p)^{n-t} \frac{p(\mathbf{X}|p)}{p(\mathbf{T}(\mathbf{x})|p)}$$

$$\begin{aligned} q(\mathbf{T}(\mathbf{x})|p) &= \frac{p^{x_1 + \cdots + x_n} (1 - p)^{n - (x_1 + \cdots + x_n)}}{\binom{n}{x_1 + \cdots + x_n} p^{x_1 + \cdots + x_n} (1 - p)^{n - (x_1 + \cdots + x_n)}} \\ &= \frac{1}{\binom{n}{T(x)}} \end{aligned}$$

does not depend on θ

例题 5.7 Example 6.2.4 Let X_1, X_2, \dots, X_n be i.i.d. $N(\mu, \sigma^2)$, where σ is unknown. $T(\mathbf{X}) = (X_1 +$

$X_2 + \cdots + X_n)/n$ is sufficient for μ

$$\begin{aligned}
 f(\mathbf{x}|\mu) &= \prod_{i=1}^n (2\pi)^{-1/2} \sigma^{-1} \exp(-(x_i - \mu)^2 / (2\sigma^2)) \\
 &= (2\pi\sigma^2)^{-n/2} \exp \left[-\sum_{i=1}^n (x_i - \mu)^2 / (2\sigma^2) \right] \\
 &= (2\pi\sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \hat{x})^2 - \frac{n}{2\sigma^2} (\hat{x} - \mu)^2 \right]
 \end{aligned}$$

$$\bar{\mathbf{X}} \sim N(\mu, \sigma^2/n)$$

$$f_{\bar{\mathbf{X}}}(t|\mu) = (2\pi\sigma^2/n)^{-n/2} \exp \left[-\frac{n}{2\sigma^2} (t - \mu)^2 \right]$$

So

$$\frac{f(\mathbf{x}|\mu)}{f_{\bar{\mathbf{X}}}(t|\mu)} = \frac{(2\pi)^{-n/2}}{2\pi n^{-1/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

does not depend on μ

例题 5.8 Example 6.2.5(Sufficient Order Statistic) Suppose X_1, X_2, \dots, X_n are i.i.d. $f(x)$. Then

$$(X_{(1)}, X_{(2)}, \dots, X_{(n)})$$

is sufficient for $f(\cdot)$. $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ is the order statistic of X_1, X_2, \dots, X_n .

5.5.2.1 Sufficient Partion

- The sufficient can be viewed as a proper partition of the sample space.

Example. Let $X_1, X_2, X_3 \sim \text{Bernoulli}(p)$, let $T = \sum x_i$. According to different values of T , the original sample space ω is partitioned onto 4 subsets.

(x_1, x_2, x_3)		t	$p(x t)$
$(0, 0, 0)$	\longrightarrow	0	1
$(1, 0, 0)$	\longrightarrow	1	1/3
$(0, 1, 0)$	\longrightarrow	1	1/3
$(0, 0, 1)$	\longrightarrow	1	1/3
$(1, 1, 0)$	\longrightarrow	2	1/3
$(0, 0, 1)$	\longrightarrow	2	1/3
$(0, 1, 1)$	\longrightarrow	2	1/3
$(1, 1, 1)$	\longrightarrow	3	1

表 5.1

$$\Omega = \{(0, 0, 0)\} \cup \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\} \\ \cup \{(0, 1, 1), (1, 0, 1), (1, 1, 0)\} \cup \{(1, 1, 1)\}$$

$$\Omega = \{(0, 0, 0)\} \cup \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\} \\ \cup \{(0, 1, 1), (1, 0, 1), (1, 1, 0)\} \cup \{(1, 1, 1)\}$$

- In each element of the partition (each of the four subset), the conditional probability of the data does not depend on θ .
- We call such a partition as **sufficient partition**
- This partition is introduced by the statistic T . Any statistic T can introduce a partition.
- Different statistic may introduce the same partition. For example, $10 \sum x_i, (\sum x_i)^2$ introduce the same partition introduced is also sufficient.
- **T is sufficient if and only if the partition introduced is also sufficient.**

例題 5.9 Sufficient Partition: One More Example

- How about the partition induced by other statistic?

Example. Let $X_1, X_2, X_3 \sim \text{Bernoulli}(p)$, let $T = X_1 + X_2$. Then partition introduced is as following.

(x_1, x_2, x_3)		t	$p(x t)$
$(0, 0, 0)$	\longrightarrow	0	$1 - p$
$(0, 0, 1)$	\longrightarrow	0	p
$(1, 0, 0)$	\longrightarrow	1	$(1 - p)/2$
$(0, 1, 0)$	\longrightarrow	1	$(1 - p)/2$
$(0, 1, 1)$	\longrightarrow	1	$p/2$
$(1, 0, 1)$	\longrightarrow	1	$p/2$
$(1, 1, 0)$	\longrightarrow	2	$1 - p$
$(0, 1, 1)$	\longrightarrow	2	p

表 5.2

The sample space is decomposed into a 3-element partition. However, in this partition, the conditional distribution still depends on p . This is not a sufficient partition, and T is not a sufficient statistic.

How to find a sufficient statistic?

定理 5.1 (Theorem 6.2.6 The Factorization Theorem)

Let $f_X(x; \theta)$ be the density of a random sample. A statistic $T(X)$ is sufficient for θ if and only if there exist functions $g(t; \theta)$ and $h(x)$, such that for any (x, θ) ,

$$f_X(x; \theta) = g(T(X); \theta)h(x)$$



- $f_X(x; \theta)$ is the joint density for the random sample x_1, \dots, x_n ,
- The density function can be seen as a product of function about T and θ , and function about x only.
- No need to calculate the conditional distribution.

This theory is most useful in finding out sufficient statistic

Proof We prove it assuming X is discrete; the continuous case is similar.

- "Only if": Let T be sufficient. Choose $g(t; \theta) = P(T(X) = t; \theta)$ and $h(x) = P(X = x | T(X) =$

$T(x)$). Since T is sufficient, $h(x)$ does not depend on θ .

$$\begin{aligned}
 f_X(x : \theta) &= P(X = x; \theta) = P(X = x | T(x) = T(x); \theta) \\
 &= P(X = x | T(X) = T(x); \theta) P(T(X) = T(x); \theta) \\
 &= P(X = x | T(X) = T(x)) P(T(X) = T(x); \theta) \\
 &= h(x) g(T(x); \theta).
 \end{aligned}$$

- "if": suppose the factorization holds, and we want to show T is sufficient for θ . Let $A_{T(x)} = \{y; T(y) = T(x)\}$, then consider

$$\begin{aligned}
 \frac{f_X(x; \theta)}{f_T(t; \theta)} &= \frac{h(x)g(T(x); \theta)}{f_T(t; \theta)} = \frac{h(x)g(T(x); \theta)}{\sum_{u \in A_{T(x)}} h(u)g(T(u); \theta)} \\
 &= \frac{h(x)g(T(x); \theta)}{g(T(x); \theta) \sum_{u \in A_{T(x)}} h(u)} = \frac{h(x)}{\sum_{u \in A_{T(x)}} h(u)}
 \end{aligned}$$

The conditional distribution does not depend on θ , hence T is sufficient for θ .

- **Example 6.2.7** X_1, \dots, X_n iid. $N(\mu, \sigma^2)$, σ known, we have $f(x|\mu) = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2] \exp[-\frac{n}{2\sigma^2} (\bar{x} - \mu)^2]$.

Since $\exp[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2]$ does not involve μ ,

$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ is a sufficient statistic for μ .

例題 5.10 Example 6.2.8: Uniform Sufficient Statistic Let X_1, X_2, \dots, X_n be iid. observations from discrete Uniform distribution on $1, 2, \dots, \theta$.

$$f(x|\theta) = \begin{cases} 1/\theta, & x = 1, 2, \dots, \theta \\ 0, & \text{otherwise} \end{cases}$$

Thus the joint pmf of X_1, \dots, X_n is

$$f(x|\theta) = \begin{cases} \theta^{-n}, & x_i \in \{1, 2, \dots, \theta\} \text{ for } i = 1, 2, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

Let

$$f(x|\theta) = \theta^{-n} I(x)_{\{1, 2, \dots, \theta\}} = \theta^{-n} I(\max\{x_i\})_{\{\max\{x_i\} \leq \theta\}}$$

$$\begin{aligned} g(t|\theta) &= \theta^{-n}, t \leq \theta \\ &= \theta^{-n} I[t \leq \theta] \end{aligned}$$

Then

$$f(x|\theta) = g(\max_{1 \leq i \leq n} x_i | \theta) h(x)$$

$\Rightarrow T(X) = \max_{1 \leq i \leq n} \{X_i\}$ is a sufficient statistic for θ .

例題 5.11 **Example 6.2.9** $X_1, \dots, X_n \sim N(\mu, \sigma^2)$.

$$\begin{aligned} f(x|\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 - \frac{n}{2\sigma^2} (\bar{x} - \mu)^2\right\} \\ &= h(x)g(T_1(x), T_2(x)|\mu, \sigma^2) \end{aligned}$$

Here,

$$h(x) \equiv 1$$

$$g(t_1, t_2|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n-1}{2\sigma^2} t_2^2 - \frac{n-1}{2\sigma^2} (t_1 - \mu)^2\right\}$$

Hence, $T_1(x) = \bar{X}$, $T_2(x) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ are sufficient statistics.

定理 5.2 (Theorem 6.2.10)

Let $X = (X_1, X_2, \dots, X_n)$ be i.i.d. observations from a pdf or pmf $f(x; \theta)$ that belongs to an exponential family given by

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum \omega_i(\theta)t_i(x)\right)$$

where $\theta = (\theta_1, \dots, \theta_d)$, $d \leq k$. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^n t_1(X_j), \dots, \sum_{j=1}^n t_k(X_j) \right)$$

is a sufficient statistic for θ



- **Example** Let X_1, \dots, X_n be i.i.d. $\text{Gamma}(\alpha, \beta)$, then $T(\mathbf{X}) = (\sum_{j=1}^n \log X_j, \sum_{j=1}^n X_j)$ are sufficient for (α, β) .
- **Example** Let X_1, \dots, X_n be i.i.d. $\text{Uniform}(\alpha, \beta)$, $\alpha < \beta$, then $(\min_{1 \leq i \leq n} X_i, \max_{1 \leq i \leq n} X_i)$ is sufficient for (α, β) .

5.5.2.2 Minimal Sufficient Statistics(MSS)

- There are multiple sufficient statistics for one parameter
- Example: $X_1, X_2, X_3 \sim \text{Bernoulli}(p)$. For p , $\sum_{i=1}^3 X_i$, $(\sum_{i=1}^3 X_i)^2$, $(X_1 + X_2, X_3)$ are all sufficient statistics
- Which is the "best" one for us?

Recall:

- Sufficient statistics: data reduction
- Best: the sufficient statistics that maximal the data reduction
- The "best" statistics has minimal data but sufficient information. We call it the **minimal sufficient statistics**.

定理 5.3 (Minimal Sufficient Statistic)

A statistic T is called a **Minimal Sufficient Statistic** if

- T is sufficient;
- For any other sufficient statistics U , $T = g(U)$ for some function g .



- For a fixed family of distribution, many sufficient statistics exist. We need to find the sufficient statistic which achieves the maximal data reduction.
- First, any one-to-one transformation of sufficient statistic is a sufficient statistic.

例题 5.12 MSS: Example Let $X_1, X_2, X_3 \sim \text{Bernoulli}(p)$. Let $T = \sum_{i=1}^3 X_i$, $U = 2X_1 + 3X_2 + 4X_3$.

Both T and U are sufficient statistics, but U is not minimal.

- How to check the minimal sufficiency?

(x_1, x_2, x_3)	t	$p(x t)$	u	$p(x u)$
(0, 0, 0)	0	1	0	1
(1, 0, 0)	1	1/3	2	1
(0, 1, 0)	1	1/3	3	1
(0, 0, 1)	1	1/3	4	1
(0, 1, 1)	2	1/3	7	1
(1, 0, 1)	2	1/3	6	1
(1, 1, 0)	2	1/3	5	1
(1, 1, 1)	2	1	9	1

- How to find a minimal sufficient statistic?

定理 5.4 (Theorem 6.2.13c: Minimal Sufficient Statistics)

Let $f_X(x; \theta)$ be the density of a random sample X , let

$$R(x, y; \theta) = \frac{f_X(x; \theta)}{f_Y(y; \theta)}$$

For a statistic T , T is minimal sufficient if $R(x; y; \theta)$ does not depend on $\theta \Leftrightarrow T(x) = T(y)$.



- Here, x and y are two random samples with the same sample size

- Sometimes, it is hard to show the equivalence.

例题 5.13MSS:Example

- Let $X_1, \dots, X_n \sim \text{Poisson}(\theta), Y_1, \dots, Y_n \sim \text{Poisson}(\theta)$, then

$$p(x; \theta) = \frac{e^{-n\theta} \theta^{\sum x_i}}{\prod y_i!}, R(x, y; \theta) = \frac{\theta^{\sum y_i - \sum x_i}}{\prod y_i! / \prod x_i!}$$

It is independent with θ if and only if $(\sum y_i = \sum x_i)$, so $T = \sum x_i$ is a minimal sufficient statistic for θ .

- Let X_1, \dots, X_n be a random sample with Cauchy distribution. Recall for Cauchy distribution, the PDF is $f(x; \theta) = \frac{1}{\pi(1+(x-\theta)^2)}$. So, the ratio is

$$R(x, y; \theta) = \frac{f(x; \theta)}{f(y; \theta)} = \frac{\prod 1/[\pi(1+(x_i-\theta)^2)]}{\prod 1/[\pi(1+(y_i-\theta)^2)]} = \frac{\prod 1/[1+(y_i-\theta)^2]}{\prod 1/[1+(x_i-\theta)^2]}$$

The result cannot be further reduced. However, note that the final result is not affected by the order of the data. Therefore, R does not depend on θ if and only if $(x_{(1)}, \dots, x_{(n)}) = (y_{(1)}, \dots, y_{(n)})$. The sufficient statistic is $T = (X_{(1)}, \dots, X_{(n)})$.

- **Example 6.2.14** X_1, \dots, X_n be i.i.d. $N(\mu, \sigma^2)$. Then $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ and $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ are minimal sufficient for μ, σ^2 .

- **Example 6.2.14** X_1, \dots, X_n be i.i.d. $Uniform(\theta, \theta + 1)$. Then

$$f_x(x|\theta) = \begin{cases} 1, & \max_i x_i - 1 < \theta < \min_i x_i \\ 0, & \text{otherwise} \end{cases}$$

This implies (Theorem 6.2.13) that $(\max_i X_i, \min_i X_i)$ is minimal sufficient for θ .

Remark 1 The above is an example of two-dimensional minimal sufficient statistic for one-dimensional parameter.

Remark 2 Any one-to-one function of minimal sufficient statistic is also a minimal sufficient statistic.

5.5.3 Ancillary Statistics

- Sufficient statistic: the statistics that contain all information about θ .
- Ancillary statistics: the statistics does not depend on θ .

定义 5.4 (Definition 6.2.16 Ancillary Statistics)

A statistic $S(X)$ of a random sample whose distribution does not depend on θ is called an ancillary statistic.



Example Let X_1, \dots, X_n be i.i.d. $Uniform(\theta, \theta + 1)$, we see that (from Example 6.1.15) $X_{(n)}, X_{(1)}$ are minimal sufficient for θ . Therefore $(X_{(n)} - X_{(1)}, \frac{X_{(n)} + X_{(1)}}{2})$ are minimal sufficient for θ . But Example 6.1.17 shows that $X_{(n)} - X_{(1)}$ is ancillary for θ .

Remark An ancillary statistic by itself may contain no information on θ , but when combine with other statistics, it may offer very important information. It is certainly not true that ancillary statistics are independent of minimal sufficient statistics.

● **Example 6.2.17** X_1, \dots, X_n are i.i.d. $Uniform(\theta, \theta + 1)$, $-\infty < \theta < \infty$. Then $R = X_{(n)} - X_{(1)}$ is ancillary.

Answer The joint pdf of $(X_{(n)}, X_{(1)})$ is

$$g(x_{(1)}, x_{(n)} | \theta) = n(n-1)(x_{(n)} - x_{(1)})^{n-2}, \theta < x_{(1)} < x_{(n)} < \theta + 1.$$

Let

$$\begin{cases} R = X_{(n)} - X_{(1)}, \\ M = (X_{(1)} + X_{(n)})/2, \end{cases}$$

then

$$f_{R,M}(r, m) = n(n-1)r^{n-2}, 0 < r < 1, \theta + (r/2) < m < \theta + 1 - (r/2).$$

So the marginal distribution of R is $f_R(r) = \int_{\theta+r/2}^{\theta+1-r/2} n(n-1)r^{n-2} dm = n(n-1)r^{n-2}(1-r), 0 < r < 1.$

\Rightarrow The pdf of R does not depend on θ . So R is ancillary for θ .

● **Example 6.2.18 (Location Family Ancillary Statistic)**

X_1, X_2, \dots, X_n are i.i.d. with cdf $F(x - \theta), -\infty < \theta < \infty$. F is a known distribution function. In this case $R = X_{(n)} - X_{(1)}$ is ancillary for θ .

● **Example 6.2.19 (Scale Family Ancillary Statistic)**

Let X_1, X_2, \dots, X_n be i.i.d. from $F(x/\sigma), \sigma > 0$. Then any statistic that depends on the sample through the $n-1$ values $X_1/X_n, \dots, X_{n-1}/X_n$ is an ancillary statistic. For example, $(X_1 + \dots + X_n)/X_n$ is ancillary. The joint distribution of $X_1/X_n, \dots, X_{n-1}/X_n$

are

$$\begin{aligned}
 F(y_1, \dots, y_{n-1} | \sigma) &= Pr\{X_1/X_n \leq y_1, \dots, X_{n-1}/X_n \leq y_{n-1}\} \\
 &= Pr\left\{\frac{\sigma Z_1}{\sigma Z_n} \leq y_1, \dots, \frac{\sigma Z_{n-1}}{\sigma Z_n} \leq y_{n-1}\right\} \\
 &= Pr\{Z_1/Z_n \leq y_1, \dots, Z_{n-1}/Z_n \leq y_{n-1}\}
 \end{aligned}$$

does not depend on σ . Z_1, \dots, Z_n are i.i.d. from $F(x)$.

Remark Ancillary statistic may still useful in estimation of θ . One example is that X_1, X_2, \dots, X_n i.i.d $N(\mu; \sigma^2)$ with σ^2 unknown. Then $T_1(X) = \frac{1}{n}(X_1 + \dots + X_n)$ is minimal sufficient for μ . But the variance estimate of $T_1(X)$ depends on S_n^2 , which is ancillary for μ .

5.5.4 Complete Statistics

定义 5.5 (Definition 6.2.21 : Complete Statistics)

] Let X be a random sample with density $f_X(x; \theta)$ and T a statistic with density $f_T(t; \theta)$. The collection of densities f_X is called complete if

$$E_\theta[g(T)] = 0 \Rightarrow P_\theta[g(T) = 0] = 1 \quad g : T \rightarrow \mathbb{R}, \theta \in \Theta.$$

T is called a Complete Statistics.



Remark:

- g is a fixed function. Say, $g(x) = x$. There is no randomness for g . The randomness of $g(T)$ comes from T .
- g does not depend on θ .
- g : a function so that $E_\theta[g(T)] = 0$ for any $\theta \in \Theta$. For any g satisfying such condition, $g(T) = 0$ with probability 1 for any θ .
- The statistic is the statistic which ensures θ is identifiable

例題 5.14 Complete Statistics: Example Example. Let $X_1, X_2, X_3 \sim \text{Bernoulli}(p)$, $\theta \in (0, 1)$. Prove that $T = \sum X_i$ is complete.

Proof: Suppose that $T \sim \text{Bernoulli}(n, \theta)$, $\theta \in (0, 1)$ and g be such that $E_\theta[g(T)] = 0$. Then we must

have

$$\begin{aligned}
 0 &= {}_\theta [g(T)] = \sum_{t=0}^n g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} \\
 &= (1-\theta) \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t \\
 &= (1-\theta) \sum_{t=0}^n g(t) \binom{n}{t} r^t
 \end{aligned}$$

where $r = \theta/(1-\theta)$. Let r be a very small number so that $g(0) \binom{n}{0} r^0$ term be the giant component, then since the summation is 0, obviously $g(0) = 0$. Similarly, we show that $g(t) = 0$ for each $t \in \{0, \dots, n\}$ must hold. Hence, T is complete.

例題 5.15 Example 6.2.23 Let $X_i \sim \text{Unif}(0, \theta)$, $i \in 1, \dots, n$, for $\theta > 0$. Recall that $T = X_{(n)}$ (the maximum of the sample) is sufficient for θ . Now, we want to prove that T is also complete.

Proof. The CDF of t is

$$F_T(t) = P(T \leq t) = P(\max X_1, \dots, X_n \leq t) = \left(\frac{t}{\theta}\right)^\theta$$

so the PDF of t is the derivative of F_t , which is $\frac{nt^{n-1}}{\theta^n}$, $0 < t < \theta$. Suppose that $g(t)$ satisfies that ${}_\theta [g(T)] = 0$, then $\int_0^\theta g(t) \frac{nt^{n-1}}{\theta^n} dt = 0$. Since it stands for all θ , the derivative of ${}_\theta [g(T)]$ also equals

to 0.

$$0 = \frac{d}{d\theta} \int_0^\theta g(t) \frac{nt^{n-1}}{\theta^n} dt =$$

$$\frac{d}{d\theta} (\theta^{-n}) \int_0^\theta g(t) nt^{n-1} dt + \frac{d}{d\theta} \left(\int_0^\theta g(t) nt^{n-1} dt \right) (\theta^{-n})$$

The first part equals to 0, since $\frac{d}{d\theta} \int_0^\theta g(t) \frac{nt^{n-1}}{\theta^n} dt = 0$. So we have

$$0 = \frac{d}{d\theta} \left(\int_0^\theta g(t) nt^{n-1} dt \right) (\theta^{-n}) = g(\theta) n \theta^{n-1}.$$

So, $g(\theta) = 0$ for any $\theta > 0$, which means that $g(x) = 0$ when $x > 0$. Recall that $T > 0$ with probability 1, so $g(T) = 0$ with probability 1, for any θ .

- **Theorem 6.2.24 (Basu's Theorem)** If $T(X)$ is a complete and minimal sufficient statistic, then $T(X)$ is independent of every ancillary statistic.
- **Theorem 6.2.25 (Complete Statistics in the exponential Family)** Let X_1, X_2, \dots, X_n be i.i.d. observations from an exponential family with pdf or pmf of the form

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^k \omega_j(\theta)t_j(x)\right)$$

where $\theta = (\theta_1, \dots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = (t_1^k(X_1), \dots, t_k^k(X_k))$$

is complete as long as the parameter space Θ contains an open set in R^k .

- **Theorem 6.2.28** If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.
- **Example 6.2.26 (Using Basu's Theorem I)** To show $g(\mathbf{X}) = \frac{X_n}{X_1 + \dots + X_n}$ and $T(\mathbf{X}) = X_1 + \dots + X_n$ are independent when X_1, X_2, \dots, X_n are i.i.d. $\exp(\text{mean}=\theta)$.
- **Example 6.2.27 (Using Basu's Theorem II)** To show \bar{X}_n and S^2 are independent if X_1, X_2, \dots, X_n are i.i.d. $N(\mu, \sigma^2)$.

5.5.5 Sufficient statistics, ancillary statistics, and complete statistics

- Sufficient statistics, ancillary statistics, and complete statistics are the statistics for data reduction.
- In past days, when the space is not enough
 - Sufficient statistics is to reduce data so that estimation through likelihood is doable.

- Ancillary statistics is to figure out the part that not related to θ
- Complete statistics is to make sure that θ is identifiable (no two θ with exactly the same model)
- Reduce the samples to be only these statistics
- Currently, thanks to the technology development, saving the data is not that difficult. These statistics are used to help understand the model and the data and accelerate the algorithm.
- Complete statistics and ancillary statistics are not popular now.

第 6 讲 Point estimation

6.1 Estimators

定义 6.1 (Definition 7.1.1)

Let $X_1, \dots, X_n \sim f(x; \theta)$. An estimator

$$\hat{\theta} = \hat{\theta}_\theta = W(X_1, \dots, X_n)$$

is a function of the data.



An estimator is a statistic. It is a random variable.

When we have an observation x_1, x_2, \dots, x_n , the corresponding result $w(x_1, x_2, \dots, x_n)$ is called an estimate of θ .

Note: estimator $\hat{\theta}$ is a random variable, and estimate is a realization of this random variable

Multiple estimator for θ .

6.2 Method of Moments(MoM)

Suppose the unknown parameter $\theta = (\theta_1, \theta_2, \dots, \theta_k)^\top$. Define a sequence of the moment functions

Data moment	Theoretical moment
$m_1 = \frac{1}{n} \sum_{i=1}^n X_i$	$\mu_1(\theta) = E[X_i]$
$m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$	$\mu_2(\theta) = E[X_i^2]$
	\vdots
$m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$	$\mu_k(\theta) = E[X_i^k]$

Let the estimator $\hat{\theta}$ satisfies that

$$m_j = \mu_j(\hat{\theta}), j = 1, \dots, k.$$

Then there are k equations. Solve the k equations to get $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$, where each of them is a function of m_1, m_2, \dots, m_k , the moments from data.

例题 6.1Example 7.2.1 Let $X_i \sim N(\mu, \sigma^2)$. Find the estimators of μ and σ^2 with MOM method.

Note that the theoretical moments are

$$E[X] = \mu, \quad E[X^2] = \mu^2 + \sigma^2.$$

The MOM estimator should satisfy that

$$\frac{1}{n} \sum_{i=1}^n X_i = \hat{\mu}, \quad \sum_{i=1}^n X_i^2 = \hat{\mu}^2 + \hat{\sigma}^2 \circ \mathcal{E}$$

Therefore, the estimators are

$$\hat{\mu} = \bar{X}_n \quad \hat{\sigma}^2 = \sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

例题 6.2 Example 7.2.2

Let $X_i \sim \text{Binomial}(k, p)$ random variable with $\theta = (k, p)$. Find the estimators of μ and p with MOM

Note that the theoretical moments are

$$E[X] = kp, \quad E[X^2] = (kp)^2 + kp(1-p).$$

The MOM estimator should satisfy that

$$\frac{1}{n} \sum_{i=1}^n X_i = \hat{k}\hat{p}, \quad \sum_{i=1}^n X_i^2 = (\hat{k}\hat{p})^2 + \hat{k}\hat{p}(1-\hat{p}) \circ \mathcal{E}$$

Therefore, the estimators are

$$\hat{p} = \bar{X}_n / \hat{k} \quad \hat{k} = \frac{\bar{X}_n^2}{\bar{X}_n^2 - \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

- Easy to calculate
- May not fit the parameter set (\hat{k} may not be an integer here)

6.3 Maximum Likelihood Estimator(MLE)

- Consider a sample $X_1, \dots, X_n \sim f(x; \theta), \theta \in \Theta$
- Observe x_1, \dots, x_n
- Which θ maximizes the possibility that we have this observation?

$$\hat{\theta} = \arg \max_{\theta \in \Theta} f_{\theta}(x_1, \dots, x_n)$$

- Similar as what did for I.I.D. Normal Model, we find the joint density:

$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta)$$

which is also a function of θ . We call $L(\theta)$ as the likelihood function.

- The estimate is called the Maximum Likelihood Estimate.

定义 6.2 (Definition 7.2.4)

For each sample point \mathbf{x} , let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ , which \mathbf{x} held fixed. Then $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimator of θ .



- Sometime maximizing $l(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x})$ is much easier than maximizing $L(\theta|\mathbf{x})$
- Remark X_1, \dots, X_n in most cases do not have to be identically distributed.

The procedure to figure out the maximiser of $L(\theta)$:

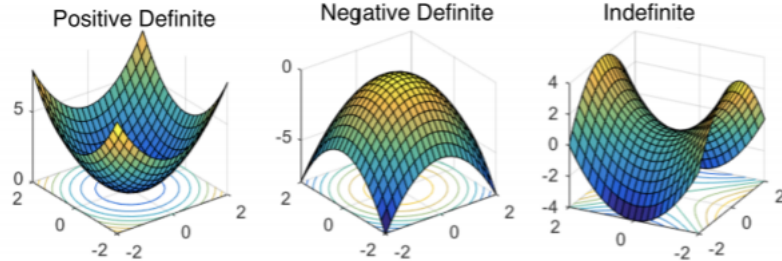
- Find the likelihood function $L(\theta)$ (actually, the joint density function)
- Find the log-likelihood function $l(\theta) = \log L(\theta)$
- Compute the gradient vector of $l(\theta)$ with respect to θ , denoted by $\nabla l(\theta)(x_1, \dots, x_n; \theta)$. For univariate case, it is the gradient only.
- Solve $\nabla l(\theta)(x_1, \dots, x_n; \theta) = 0$, with respect to $\theta \in \Theta$, call this solution $\tilde{\theta}_n$, check whether $H(\tilde{\theta}_n)$

is negative definite. If yes, then $\tilde{\theta}_n = \hat{\theta}_n$. Here, $H(\theta)$ is the Hessian matrix for $l(\theta)$, defined as

$$\begin{pmatrix} \frac{\partial(\theta)}{\partial\theta_1^2} & \frac{\partial(\theta)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial(\theta)}{\partial\theta_1\partial\theta_k} \\ \frac{\partial(\theta)}{\partial\theta_2\partial\theta_1} & \frac{\partial(\theta)}{\partial\theta_2^2} & \cdots & \frac{\partial(\theta)}{\partial\theta_2\partial\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial(\theta)}{\partial\theta_k\partial\theta_1} & \frac{\partial(\theta)}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial(\theta)}{\partial\theta_k^2} \end{pmatrix}$$

Hessian Matrix ■ When $k = 1$, the Hessian Matrix is the second derivative of $l(\theta)$.

- The matrix can be seen as the second derivative for a multivariate function
 - Negative definite indicates local maximum
 - Since there is only one solution, the local maximum is also the global maximum (if Θ is properly defined)
- It may be hard to figure out the solution for $\nabla l(\theta) = 0$ analytically. Computationally, some methods can be applied, such as Newton's method. Details in applied mathematics/computing mathematics.
 - The general Hessian matrix $H(\theta)$ may be complicated, yet $H(\hat{\theta}_n)$ is usually easier since $\hat{\theta}_n$



satisfies $\nabla l(\hat{\theta}_n) = 0$.

- If there are multiple solutions, then we should check the values to figure out the maximal one, check the Hessian matrix at this maximal solution, and also check $l(\theta)$ when some θ goes to infinity.
- If Θ is not properly defined, it is possible that the maximal is achieved at the bound of Θ , not the maximum point. Or, the solution $\tilde{\theta}_n$ does not exist in Θ .

例题 6.3MLE:Example1 Let $X_1, X_2, \dots, X_n \sim \text{Poisson}(\lambda)$. Find the MLE of λ , given observations

x_1, \dots, x_n .

The likelihood function is

$$L(\lambda) = f(x_1, \dots, x_n; \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \frac{1}{\prod_{i=1}^n x_i!}$$

The log-likelihood function is

$$l(\lambda) = \log L(\lambda) = -n\lambda + \left(\sum_{i=1}^n x_i\right) \log \lambda - \sum_{i=1}^n \log x_i!$$

The gradient vector is a derivative:

$$\frac{d}{d\lambda} l(\lambda) = -n + \frac{1}{\lambda} \left(\sum_{i=1}^n x_i\right)$$

Set it to be 0, and solve the equation, we have $\tilde{\lambda}_n = \bar{x}_n$. Note that

$$\frac{d^2}{d\lambda^2} l(\lambda) \big|_{\lambda=\tilde{\lambda}} = -\frac{1}{\tilde{\lambda}^2} \left(\sum_{i=1}^n x_i\right) \big|_{\lambda=\tilde{\lambda}} = -n/\bar{X}_n < 0$$

So, the MLE is $\hat{\lambda}_n = \bar{X}_n$.

例题 6.4MLE:Example 2 Let $X_1, X_2, \dots, X_n \sim \text{Exp}(\lambda)$, Find the MLE of λ , given observations x_1, x_2, \dots, x_n .

The likelihood function is

$$L(\lambda) = f(x_1, \dots, x_n; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = e^{-\lambda \sum_{i=1}^n x_i} \lambda^n$$

The log-likelihood function is

$$l(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \left(\sum_{i=1}^n x_i \right)$$

The gradient vetor is a derivative:


$$\frac{d}{d\lambda} l(\lambda) = n/\lambda - \sum_{i=1}^n x_i.$$

Set it to be 0, and solve the equation, we have $\tilde{\lambda}_n = 1/\bar{x}_n$. Check thte second derivative at $\tilde{\lambda}_n = \bar{x}_n$. Note that

$$\frac{d^2}{d\lambda^2} l(\lambda)|_{\lambda=\tilde{\lambda}} = -\frac{n}{\tilde{\lambda}^2}|_{\lambda=\tilde{\lambda}} < 0.$$

So, the MLE is $\hat{\lambda}_n = 1/\bar{X}_n$.

定理 6.1 (Invariance of MLE)

Let $\hat{\theta}_n$ be the MLE of θ . Then for any function $\tau : \Theta \rightarrow \mathbb{R}^k$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta}_n)$. 

- For a function of θ , we do not need to calculate the derivative with respect to this new function, but introduce in $\hat{\theta}_n$ onto the function directly.
- Introduce the method as [Profile likelihood](#). Assuming we are interested in only part of the parameters η , where $\theta = (\eta, \xi)$. Then the profile likelihood is defined by

$$L(\eta) = \max_{\xi} L(\eta, \xi)$$

Maximizing $L(\eta)$ gives the same MLE for $\hat{\eta}_n$ with the one we get from $\hat{\theta} = (\hat{\eta}_n, \hat{\xi}_n)$.

6.4 Bayes Estimators

Let $X \sim f(\mathbf{x}|\theta)$, $\theta \sim \pi(\theta)$, where $\pi(\theta)$ is the prior distribution of θ . Then after observing $X = \mathbf{x}$, the posterior distribution of θ is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

where $m(\mathbf{x})$ is the marginal distribution of \mathbf{X} ,

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta.$$

例题 6.5(Example 7.2.14) Let X_1, X_2, \dots, X_n be i.i.d. Bernoulli(p). Then $Y = \sum_{i=1}^n X_i$ is Bin(n, p).

We assume the prior distribution on p is Beta(α, β). Then

$$\begin{aligned} f(y, p) &= \left[\binom{n}{y} p^y (1-p)^{n-y} \right] \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \right] \\ &= \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1} (1-p)^{n-y+\beta-1} \\ f(y) &= \int_0^1 f(y, p) dp = \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(y + \alpha)\Gamma(n - y + \beta)}{\Gamma(n + \alpha + \beta)} \end{aligned}$$

Therefore $f(p|y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1} (1-p)^{n-y+\beta-1}$, which is Beta($y + \alpha, n - y + \beta$). Bayes estimator of p is the mean of $f(p|y)$, which is

$$\hat{p}_B = \frac{y + \alpha}{\alpha + \beta + n}.$$

定义 6.3 (Definition 7.2.6 (Conjugate Family))

Let \mathcal{F} denote the class of pdf or pmf $f(x|\theta)$ (indexed by θ). A class Π of prior distribution is a conjugate family of \mathcal{F} if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all prior in Π , and all $x \in \mathcal{X}$.



$$P(\theta) = \pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int_{\theta} f(\mathbf{y}|\theta)\pi(\theta)d\theta} \quad (12.3')$$

- Estimating the normalizing constant
- Markov chain Monte Carlo (MCMC): Monte Carlo integration and Markov chain sampling

we estimated unknown parameters using the methods:

- Maximum likelihood : Newton–Raphson(NR)
- MCMC
- NR 比 MCMC 收敛的快。
- 对于多峰分布, NR 只能找到一个, 而 MCMC 可以找到多个。
- NW 对于固定的似然初始值, 它的路径是一样的; 而 MCMC 即使初始值相同, 它的路径也是随机的。

- **Target density** $P(\theta)$ is not always achievable because it may have a complex, or even unknown, form.
- Markov chains provide a method of drawing samples from target densities (regardless of their complexity).
- Using these conditional steps, we build up a chain of samples $(\theta^1, \dots, \theta^M)$ after specifying a starting value θ^0

Markov property

$$P(\theta^i = a | \theta^{i-1}, \theta^{i-2}, \dots, \theta^0) = P(\theta^i = a | \theta^{i-1})$$

An algorithm for creating a Markov chain for a target probability density $P(\theta)$ is:

- 1 Choose an initial value θ^0 . The restriction on the initial value is that it needs to be within the distribution of $P(\cdot)$, so that $P(\theta) > 0$
- 2 Create a new sample using $\theta^1 \sim \pi(\theta^1 | \theta^0, \mathbf{y})$
- 3 Repeat step 2 M times, each time increasing both indices by 1.

Transitional density: $\pi(\theta^{i+1} | \theta^i)$, Normal distribution.

6.4.1 The Metropolis–Hastings sampler

- The Metropolis–Hastings sampler works by randomly **proposing a new value** θ^*
- If this proposed value is **accepted** (according to a criterion below), $\theta^{i+1} = \theta^*$
- If this proposed value is **rejected** (according to a criterion below), $\theta^{i+1} = \theta^i$
- Another proposal is made and the chain progresses by assessing this new proposal
- $\theta^* = \theta^i + Q$, Q is called the **proposal density**, $N(0,1)$ or $U[-1,1]$
- The acceptance criterion is:

$$\theta^{i+1} = \begin{cases} \theta^* & \text{if } U < \alpha \\ \theta^i & \text{otherwise} \end{cases}$$

where $U \sim U[-1, 1]$ and

$$\alpha = \min\left\{\frac{\pi(\theta^*|\mathbf{y})}{\pi(\theta^i|\mathbf{y})} \cdot \frac{Q(\theta^i|\theta^*)}{Q(\theta^*|\theta^i)}, 1\right\}$$

where $P(\theta|\mathbf{y})$ is the probability of θ given the data \mathbf{y} (the likelihood)

If the proposal density is symmetric ($Q(a|b) = Q(b|a)$), then

$$\alpha = \min\left\{\frac{\pi(\theta^*|\mathbf{y})}{\pi(\theta^i|\mathbf{y})}, 1\right\}$$

例题 6.6 The Metropolis–Hastings sampler 设 $Y_1, Y_2, \dots, Y_n \sim^{iid} N(\mu, \sigma^2)$, (μ, σ^2) 的先验分布为 $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$ 。计算 $E(\mu|Y = y)$ 和 $E(\sigma^2|Y = y)$ 。

解: (μ, σ^2) 后验分布为

$$\pi(\mu, \sigma^2|Y = y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\}$$

- 《统计模拟及其 R 实现》 P203
- 为了比较方法, 假设 $Y_1, Y_2, \dots, Y_n \sim^{iid} N(2, 4^2)$

M-H Rcode

```

1 y=rnorm(100,2,4)
2 #####Metropolis-Hastings
3 exam8.5M=function(y,n,a,b){
4   th=matrix(0,ncol=2,nrow=n)
5   m=length(y)
6   th[1,]=c(a,b)
7   pth=matrix(0,ncol=2,nrow=n)
8   accept=NULL

```

```

9   for(i in 1:(n-1)){
10     Q<-runif(2,-0.2,0.2) # tryU[-1,1]
11     pth[i+1,1]<-th[i,1]+Q[1]
12     pth[i+1,2]<-th[i,2]+Q[2]
13     U<-runif(1,0,1)
14     aa=((th[i,2]/pth[i+1,2])^(m+2))*exp(-(sum((y-ptth[i+1,1])^2))/(2*(pth[i+1,2])^2)+
15       (sum((y-th[i,1])^2))/(2*(th[i,2])^2))
16     alpha<-min(aa,1)
17     if (U>=alpha){th[i+1,1]=th[i,1]
18       th[i+1,2]=th[i,2]
19       accept[i]=0}
20     if (U<alpha){th[i+1,1]= pth[i+1,1]
21       th[i+1,2]= pth[i+1,2]
22       accept[i]=1}
23   }
24   list(mu=th[,1],sig=th[,2],accept=accept)

```

```
25 }  
26  
27 fitm=exam8.5M(y,5000,1,1)  
28 par(mfrow=c(1,2))  
29 plot(fitm$mu)  
30 plot(fitm$sig)  
31  
32 mean(fitm$mu[1000:5000])  
33 mean(fitm$sig[1000:5000])  
34 mean(y)  
35 sd(y)  
36 sum(fitm$accept)
```

6.4.2 The Gibbs sampler

- The Gibbs sampler is another way of generating a Markov chain.
- It splits the parameters into a number of components and then updates each one in turn.

- For the beetle mortality example, a Gibbs sampler to update the two unknown parameters would be:

- 1 Assign an initial value to the two unknowns: β_1^0 and β_2^0
- 2 (a) Generate $\beta_2^1 \sim \pi(\beta_2 | \mathbf{y}, \beta_1^0)$
 (b) Generate $\beta_1^1 \sim \pi(\beta_1 | \mathbf{y}, \beta_2^0)$
- 3 Repeat the step 2 M times, each time increasing the sample indices by 1.

上面举例比较简单，复杂的见《统计模拟及其 R 实现》P199 例 8.4

例题 6.7 The Gibbs sampler 设 $Y_1, Y_2, \dots, Y_n \sim iid N(\mu, \sigma^2)$, (μ, σ^2) 的先验分布为 $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$ 。计算 $E(\mu | Y = y)$ 和 $E(\sigma^2 | Y = y)$ 。

解： (μ, σ^2) 后验分布为

$$\pi(\mu, \sigma^2 | Y = y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\}$$

则：

$$\pi(\mu | \sigma^2, y) \propto \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\} \propto N\left(\bar{y}, \frac{\sigma^2}{n}\right) \quad (1)$$

$$\pi(\sigma^2|\mu, y) \propto \left(\frac{1}{\sigma^2}\right) \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\} \propto IG\left(\frac{n}{2}, \frac{\sum_{i=1}^n (y_i - \mu)^2}{2}\right) \quad (2)$$

例：考虑 54 个老人智力得分。注：中科大张伟平《计算统计讲义》

Table 7.8 *Symptoms of senility (s=1 if symptoms are present and s=0 otherwise) and WAIS scores (x) for N=54 people.*

x	s	x	s	x	s	x	s	x	s
9	1	7	1	7	0	17	0	13	0
13	1	5	1	16	0	14	0	13	0
6	1	14	1	9	0	19	0	9	0
8	1	13	0	9	0	9	0	15	0
10	1	16	0	11	0	11	0	10	0
4	1	10	0	13	0	14	0	11	0
14	1	12	0	15	0	10	0	12	0
8	1	11	0	13	0	16	0	4	0
11	1	14	0	10	0	10	0	14	0
7	1	15	0	11	0	16	0	20	0
9	1	18	0	6	0	14	0		

考虑 Logit 模型：

$$Y_i \sim \text{Bin}(1, \pi_i), \quad \log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x_i, \quad i = 1, 2, \dots, 54$$

则似然函数为：

$$\begin{aligned} f(\mathbf{y}|\beta_0, \beta_1) &= \prod_{i=1}^n \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-y_i} \\ &= \exp \left\{ \sum_{i=1}^n [(\beta_0 + \beta_1 x_i) y_i - \log(1 + e^{\beta_0 + \beta_1 x_i})] \right\} \end{aligned}$$

考虑 β_0, β_1 的先验分布为独立的正态分布：

$$\beta_j \sim N(\mu_j, \sigma_j^2)$$

从而后验分布为：

$$\begin{aligned} f(\beta_0, \beta_1|\mathbf{y}) &\propto f(\mathbf{y}|\beta_0, \beta_1) \pi(\beta_0, \beta_1) \\ &\propto \exp \left\{ \sum_{i=1}^n [(\beta_0 + \beta_1 x_i) y_i - \log(1 + e^{\beta_0 + \beta_1 x_i})] \right. \\ &\quad \left. - \frac{(\beta_0 - \mu_0)^2}{\sigma_0^2} - \frac{(\beta_1 - \mu_1)^2}{\sigma_1^2} \right\} \end{aligned}$$

6.5 EM Algorithm

- Consider a mixture model:

$$Z_i \sim \text{Bernoulli}(\theta), \quad X_i|Z_i \sim N(Z_i, \sigma^2).$$

However, our observations are X_i 's only.

- Since Z_i 's are missing. We call it as **missing data** problem
- The MLE is difficult to solve analytically for this problem.
- One generally used procedure is EM algorithm.
- Now there are two unknown things: the missing data z and the parameters θ
- If we know z , the MLE is found by

$$\arg \max_{\theta} \log f(x, z|\theta).$$

- If we know θ , then we have the conditional distribution for z , which is

$$f(z|x, \theta) = \frac{f(x, z|\theta)}{f(x|\theta)} = \frac{f(x, z|\theta)}{\int f(x, z|\theta) dz}$$

- Combine these two things, find θ as

$$\arg \max_{\theta} Q(\theta, \theta') = \arg \max_{\theta} E_Z \left[\log f(x, z|\theta)|x; \theta' \right]$$

where the expectation is taken with respect to the conditional distribution of Z , $f(z|x, \theta')$.

1. Set initial value $\theta^0, N = 1$.
2. Expectation step:
 - With θ^{N-1} , find $f(z|x; \theta^{N-1})$
 - Compute $Q(\theta, \theta^{N-1}) = E_Z \left[\log f(x, z; \theta)|x; \theta' \right]$, where $Z \sim f(z|x; \theta')$.
3. Maximization step: Find θ^N as

$$\theta^N = \arg \max_{\theta} Q(\theta, \theta^{N-1})$$

4. Repeat Steps 2-3 until $\|\theta^{N-1} - \theta^N\| \leq \epsilon$, where ϵ is a pre-set threshold. Or, stop the algorithm when N is large enough.

Remark:

- Very popular method, since very good for complicated models
- Seems different in different applications
- However, maybe trapped by local maxima

例題 6.8 The EM Algorithm: Example Suppose we observe $Z_{1:n}$ and $Y_{1:n}$ both independent random variables and independent of each other. In particular, $Y_i \sim \text{Poisson}(\tau_i)$ and $Z_i \sim \text{Poisson}(\tau_i)$, where $\theta = (\beta, \tau_1, \dots, \tau_n) \in R_+^{n+1}$ are the parameters.

- If we have the **full data**, then the joint density is

$$f_Y(y|\theta) = \prod_{i=1}^n \frac{(\beta\tau_i)^{y_i}}{y_i!} e^{-\beta\tau_i} \quad f_Z(z|\theta) = \prod_{i=1}^n \frac{\tau_i^{z_i}}{z_i!} e^{-\tau_i}$$

It is straightforward to find the MLEs;

$$\hat{\beta}_n = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n \hat{z}_i}, \quad \hat{\tau}_i = \frac{y_i + z_i}{\hat{\beta}_n + 1}, i = 1, \dots, n.$$

- Now, if z_1 was **missing**, we have the marginal data likelihood of the observations:

$$\begin{aligned} f(y, z_{2:n}; \theta) &= \frac{(\beta\tau_1)^{y_1}}{y_1!} e^{-\beta\tau_1} \left(\prod_{i=2}^n \frac{(\beta\tau_i)^{y_i}}{y_i!} e^{-\beta\tau_i} \frac{\tau_i^{z_i}}{z_i!} e^{-\tau_i} \right) \sum_{z_1=0}^{\infty} \frac{\tau_1^{z_1}}{z_1!} e^{-\tau_1} \\ &= \frac{(\beta\tau_1)^{y_1}}{y_1!} e^{-\beta\tau_1} \left(\prod_{i=2}^n \frac{(\beta\tau_i)^{y_i}}{y_i!} e^{-\beta\tau_i} \frac{\tau_i^{z_i}}{z_i!} e^{-\tau_i} \right) \end{aligned}$$

Now the Q function is

$$\begin{aligned}
Q(\theta, \theta') &= E_{Z_1} \left[\log f(y, z | \theta) \right] \\
&= \sum_{z_1=0}^{\infty} \log \left[\prod_{i=1}^n \frac{(\beta \tau_i)^{y_i}}{y_i!} e^{-\beta \tau_i} \frac{\tau_i^{z_i}}{z_i!} e^{-\tau_i} \right] \frac{(\tau'_1)^{z_1}}{z_1!} e^{-\tau'_1} \\
&= \sum_{i=1}^n \left(-\beta \tau_i + y_i [\log \beta + \log \tau_i] - \log y_i \right) + \sum_{i=2}^n \left[-\tau_i + z_i \log \tau_i - \log z_i! \right] \\
&\quad + \sum_{z_1=0}^{\infty} \log \left(-\tau_1 + z_1 \log \tau_1 - \log z_1! \right) \frac{(\tau'_1)^{z_1}}{z_1!} e^{-\tau'_1} \\
&= \sum_{i=1}^n \left(-\beta \tau_i + y_i [\log \beta + \log \tau_i] \right) + \sum_{i=2}^n \left[-\tau_i + z_i \log \tau_i \right] \\
&\quad + \sum_{z_1=0}^{\infty} \log \left(-\tau_1 + z_1 \log \tau_1 - \log z_1! \right) \frac{(\tau'_1)^{z_1}}{z_1!} e^{-\tau'_1} + \sum_{i=1}^n -\log y_i \\
&\quad + \sum_{i=2}^n -\log z_i!
\end{aligned}$$

Note that

$$\sum_{z_1=0}^{\infty} \log \left(-\tau_1 + z_1 \right) \frac{(\tau'_1)^{z_1}}{z_1!} e^{-\tau'_1} = -\tau_1 + \tau'_1 \log \tau_1$$

and the last term does not depend on θ , which can be denoted as C . So we have

$$\begin{aligned} Q(\theta, \theta') \\ = \sum_{i=1}^n \left(-\beta \tau_i + y_i [\log \beta + \log \tau_i] \right) + \sum_{i=2}^n \left[-\tau_i + z_i \log \tau_i \right] - \tau_1 + \tau'_1 \log \tau_1 + C \end{aligned}$$

Maximizing $Q(\theta, \theta')$ w.r.t. θ , the solution is

$$\beta = \frac{\sum_{i=1}^n y_i}{\tau'_1 + \sum_{i=2}^n z_i} \quad \tau_1 = \frac{\tau'_1 + y_1}{\beta + 1}, \quad \tau_i = \frac{y_1 + z_i}{\beta + 1}, i = 2, \dots, n$$

6.6 Methods of Evaluating Estimators

6.6.1 Unbiasedness

定义 6.4

Say that $\hat{\theta} = W(X_1, \dots, X_n)$ is an estimator of θ , then it would be good if it satisfies that

$$E[\hat{\theta}] = \theta$$



Let $\hat{\theta}$ be an estimator of a parameter θ . Then the bias of $\hat{\theta}$ is defined as

$$Bias(\hat{\theta}; \theta) = E_{\theta}[\hat{\theta}] - \theta$$

If $Bias(\hat{\theta}) = 0$, then we say $\hat{\theta}$ is unbiased.

■ $E_{\theta}[\hat{\theta}]$ means the expectation of $\hat{\theta}$ when the underlying parameter equals to θ .

■ The bias is a function of θ . For unbiased estimators, the bias is a function that always equals to 0.

例题 6.9 Unbiasedness: Example ■ Let $X_1, \dots, X_n \sim \text{Exp}(\lambda)$. Estimate λ .

Recall that the MLE for exponential distribution is $1/\bar{X}_n$. Let the estimator be $\hat{\lambda} = 1/\bar{X}_n$. Note that

$n\bar{X}_n \sim \text{Gamma}(n, \lambda)$. Therefore, the bias is

$$\text{Bias}(\hat{\lambda}, \lambda) = \frac{n}{n-1}\lambda - \lambda = \frac{1}{n-1}\lambda$$

Therefore, the MLE $\hat{\lambda}$ is a biased estimator. However, when $n \rightarrow \infty$, the bias is close to 0.

■ Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Find the bias for the sample variance. The sample variance is $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. The bias is

$$\text{Bias}(\hat{\sigma}^2, \sigma^2) = E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2\right] - \sigma^2 = 0$$

So, the sample variance is an unbiased estimator.

■ In the previous normal example, we show that the bias for sample variance is 0.

■ If we take the estimator as $\tilde{\sigma}^2 = \frac{n}{n-1} (X_1^2 - \bar{X}_n^2)$, then

$$E[\tilde{\sigma}^2] - \sigma^2 = \frac{n}{n-1} [E[X_1^2] - E[\bar{X}_n^2]] - \sigma^2 = \frac{n}{n-1} \times \frac{n-1}{n} \sigma^2 - \sigma^2 = 0,$$

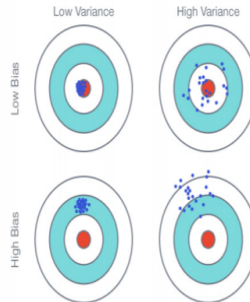
which is also unbiased.

■ Which estimator is better? the sample variance or $\tilde{\sigma}^2$? Variance

Let $\hat{\theta}$ be an estimator of the parameter θ . Then the variance of $\hat{\theta}$ is defined as

$$\text{Var}(\hat{\theta}; \theta) = \text{Var}_{\theta}(\hat{\theta}).$$

- Targeting at θ , the estimator with smaller variance is better.
- For the previous example, the variance for sample variance is $2\sigma^4/(n-1)$, but for $\tilde{\sigma}^2$ is approximately σ^4 . So, the sample variance is a better estimator.



6.6.2 Best Unbiased Estimators

定义 6.5 (Definition 7.3.7)

An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $E_{\theta} W^* = \tau(\theta)$ for all θ and, for any other estimator W with $E_{\theta} W = \tau(\theta)$, we have

$$\text{Var}_{\theta} W^* \leq \text{Var}_{\theta} W \text{ for all } \theta.$$

W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$.



Finding UMVUE is not easy.

- Does the UMVUE exist?
- Not necessarily. It is possible that UMVUE does not exist.
- How to prove one estimator is UMVUE?
- There is a lower bound for the variance of unbiased estimators. If there is one unbiased estimator with variance approaching the lower bound, then it is UMVUE.
- How to find the UMVUE?

定理 6.2 (Theorem 7.3.9 (Cramér-Rao Inequality))

Let X_1, X_2, \dots, X_n be a sample with pdf $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X})$ be any estimator satisfying

$$\frac{d}{d\theta} E_{\theta} W(\mathbf{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} [W(\mathbf{x}) f(\mathbf{x}|\theta)] d\mathbf{x}$$

and $\text{Var}_{\theta} W(\mathbf{X}) < \infty$. Then

$$\text{Var}_{\theta} W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta} E_{\theta} W(\mathbf{X}) \right)^2}{E_{\theta} \left(\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right)}$$

In particular, if $W(\mathbf{X})$ is an unbiased estimator of θ , then

$$\text{Var}_{\theta} W(\mathbf{X}) \geq \frac{1}{E_{\theta} \left(\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right)}$$



- **Corollary 7.3.10 (Cramér-Rao Inequality, iid case)** If X_1, X_2, \dots, X_n are i.i.d. $f(x|\theta)$, and

the condition of Theorem 7.3.9 are satisfied, then

$$\text{Var}_\theta (W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\mathbf{X}) \right)^2}{n E_\theta \left(\left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right)}$$

6.6.2.1 Score and Fisher Information

■ An important item here is $E_\theta \left[\left(\frac{\partial}{\partial \theta} \log(f(X_1, \dots, X_n; \theta)) \right)^2 \right]$

■ Actually, we have some notions and lemmas w.r.t. this quantity

Score function

Let X_1, \dots, X_n be with joint density $f(x_1, x_2, \dots, x_n; \theta)$. The score function is the derivative of the log-likelihood function, which is

$$S_n(\theta) = \frac{\partial}{\partial \theta} \log(f(X_1, \dots, X_n; \theta))$$

If X_1, \dots, X_n are i.i.d. with density $f(x; \theta)$, then the score function equals to

$$\frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i; \theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(X_i; \theta)$$

定理 6.3

Under regularity conditions,

$$E_{\theta}[S_n(\theta)] = 0$$



Proof. The expectation of the score function is

$$\begin{aligned} E_{\theta}[S_n(\theta)] &= \int \frac{\partial \log(f(x_1, \dots, x_n; \theta))}{\partial \theta} f(x_1, \dots, x_n; \theta) dx_1 \cdots dx_n \\ &= \int \frac{\frac{\partial}{\partial \theta} f(x_1, \dots, x_n; \theta)}{f(x_1, \dots, x_n; \theta)} f(x_1, \dots, x_n; \theta) dx_1 \cdots dx_n \\ &= \int \frac{\partial}{\partial \theta} f(x_1, \dots, x_n; \theta) dx_1 \cdots dx_n \\ &= \frac{\partial}{\partial \theta} \int f(x_1, \dots, x_n; \theta) dx_1 \cdots dx_n \\ &= \frac{\partial}{\partial \theta} 1 = 0 \end{aligned}$$

Note: If θ mismatches, it may not hold. It is possible $E_{\theta_1}[S_n(\theta_2)] \neq 0$

定义 6.6 (Fisher Information)

Let X_1, \dots, X_n be with joint density $f(x_1, x_2, \dots, x_n; \theta)$. The Fisher information is the variance of the score function, which is

$$I_n(\theta) = \text{Var}_\theta(S_n(\theta)) = E\left[\left[\frac{\partial}{\partial \theta} \log f(X_1, \dots, X_n; \theta)\right]^2\right]$$



If X_1, \dots, X_n are i.i.d. with density $f(x; \theta)$, then the Fisher information is $I_n(\theta) = nI(\theta)$ where $I(\theta)$ is the Fisher information for single observation.

■ $S_n(\theta) = 0$, yet $\text{Var}_\theta(S_n(\theta))$ is a function of θ

■ Proof. in i.i.d. case, the score function is $S_n(\theta) = \sum_{i=1}^n \log f(X_i; \theta)$

$$\begin{aligned} I_n(\theta) &= \text{Var}_\theta(S_n(\theta)) = \text{Var}\left(\sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(X_i; \theta)\right) \\ &= \sum_{i=1}^n \text{Var}\left(\frac{\partial}{\partial \theta} \log f(X_i; \theta)\right) = n \text{Var}\left(\frac{\partial}{\partial \theta} \log f(X_1; \theta)\right) = nI(\theta) \end{aligned}$$

- According to the Cramer-Rao lower bound, all the unbiased estimator for θ has variance larger than $1/I_n(\theta)$. So $I_n(\theta)$ gives us the bound for the information we can get from the data. That's

why we call it as Information.

■ Another statement of Cramer-Rao lower bound: Corollary: Unbiased Estimators

Let X_1, \dots, X_n be i.i.d. samples with density $f(x; \theta)$ and let $W : X^n \rightarrow \mathbb{R}$ be an unbiased estimator of $\tau(\theta)$. Suppose the conditions hold, then

$$\text{Var}(W; \theta) \geq \frac{\tau'(\theta)^2}{I_n(\theta)} = \frac{\tau'(\theta)^2}{nI_n(\theta)}$$

■ Obviously, the variance will converge to 0 when n increases.

■ The best unbiased estimator has convergence rate at $1/\sqrt{n}$.

定理 6.4 (Lemma 7.3.11)

$f(x|\theta)$ satisfies

$$\frac{d}{d\theta} E_{\theta} \left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right) = \int \frac{\partial}{\partial \theta} \left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right) f(x|\theta) \right] dx$$

(true for an exponential family), then

$$E_{\theta} \left(\left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right) = -E_{\theta} \left(\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right).$$



Proof. In short, we denote $X = (X_1, X_2, \dots, X_n)$. For the L.H.S, there is

$$E\left[\left(\frac{\partial}{\partial \theta} \log f(X; \theta)\right)^2\right] = E\left[\frac{1}{(f(X; \theta))^2} \left(\frac{\partial}{\partial \theta} f(X; \theta)\right)^2\right]$$

For the R.H.S, we have

$$\begin{aligned} -E\left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta)\right] &= -E\left[\frac{\partial}{\partial \theta} \frac{1}{f(X; \theta)} \frac{\partial f(X; \theta)}{\partial \theta}\right] \\ &= E\left[\frac{1}{(f(X; \theta))^2} \left(\frac{\partial}{\partial \theta} f(X; \theta)\right)^2\right] - E\left[\frac{1}{f(X; \theta)} \frac{\partial^2 f(X; \theta)}{\partial \theta^2}\right] \\ &= E\left[\frac{1}{(f(X; \theta))^2} \left(\frac{\partial}{\partial \theta} f(X; \theta)\right)^2\right] - \int \frac{\partial^2 f(X; \theta)}{\partial \theta^2} dx \\ &= L.H.S - \frac{\partial^2}{\partial \theta^2} \int f(X; \theta) dx = L.H.S \end{aligned}$$

- **Example 7.3.12** \bar{X} is UMVUE for λ if X_1, \dots, X_n are i.i.d. Poisson(λ). From [Theorem 7.3.9](#),

we have for any unbiased estimator $W(\mathbf{X})$ of λ .

$$\text{Var}_\lambda W(\mathbf{X}) \geq \frac{1}{-nE_\lambda \left[\frac{\partial^2}{\partial \lambda^2} \log f(\mathbf{x}|\lambda) \right]} \quad (6.1)$$

$$\log f(\mathbf{x}|\lambda) = \log \left[e^{-\lambda} \frac{\lambda^x}{x!} \right] = -\lambda + x \log \lambda - \log x!$$

$$\frac{\partial^2}{\partial \lambda^2} \log f(\mathbf{x}|\lambda) = -x \frac{1}{\lambda^2}.$$

$$\text{Therefore, } -E_\lambda \left[\frac{\partial^2}{\partial \lambda^2} \log f(\mathbf{x}|\lambda) \right] = \frac{1}{\lambda^2} E_\lambda X = \frac{1}{\lambda}.$$

$$(6.1) \text{ Becomes } \text{Var}_\lambda(W(\mathbf{X})) \geq \frac{\lambda}{n}.$$

$$\text{But } \text{Var}_\lambda(\bar{X}) = \frac{\lambda}{n}.$$

- **Example 7.3.13 (Unbiased Estimator for Scale Parameter)** Let X_1, \dots, X_n be i.i.d. with pdf $f(x|\theta) = \frac{1}{\theta}$, $0 < x < \theta$. Since $\frac{\partial}{\partial \lambda} \log f(x|\theta) = -\frac{1}{\theta}$, we have

$$E_\theta \left[\frac{\partial}{\partial \lambda} \log f(x|\theta) \right] = -\frac{1}{\theta^2}$$

So if W is unbiased for θ , then

$$\text{Var}_\theta(W) \geq \frac{\sigma^2}{n}.$$

- On the other hand, $Y = \max(Y_1, \dots, Y_n)$ is a sufficient statistic. $f_Y(y|\theta) = ny^{n-1}/\theta^n$, $0 < y < \theta$. So

$$E_\theta Y = \int_0^\theta y \cdot \frac{ny^{n-1}}{\theta^n} dy = \frac{n}{n+1}\theta,$$

showing that $\frac{n+1}{n}Y$ is an unbiased estimator of θ .

$$\begin{aligned} \text{Var}_\theta \left(\frac{n+1}{n}Y \right) &= \left(\frac{n+1}{n} \right)^2 \text{Var}_\theta(Y) \\ &= \left(\frac{n+1}{n} \right)^2 [E_\theta Y^2 - (EY)^2] \\ &= \left(\frac{n+1}{n} \right)^2 \left[\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta \right)^2 \right] \\ &= \frac{1}{n(n+2)}\theta^2, \end{aligned}$$

which is uniformly smaller than θ^2/n . Cramér-Rao lower bound Theorem is not applicable to

this pdf since $\frac{d}{d\theta} \int_0^\theta h(x) f(x|\theta) dx \neq \int_0^\theta h(x) \frac{\partial}{\partial \theta} f(x|\theta) dx$.

例題 6.10 Example of CRLB-Normal Example Let X_1, \dots, X_n be a random sample from the $N(\mu, \sigma^2)$ distribution. Find the CRLB and, in case 1 and 2, check whether it is equalled, for the variance of an unbiased estimator of

1. μ when σ^2 is known,
2. σ^2 when μ is known
3. μ when σ^2 is unknown
4. σ^2 when μ is unknown

Solution: The sample joint pdf is

$$f_X(X|\theta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$$

and

$$\log f_X(X|\theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2/\sigma^2$$

1. When σ^2 is known $\theta = \mu$ and

$$\log f_X(X|\theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2$$

$$S(X) = \frac{\partial}{\partial \theta} \log f_X(X|\theta) = \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2 = \frac{n}{\sigma^2} [\bar{x} - \theta]$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, is a n unbiased estimator of $\theta = \mu$ whose variance equals the CRLB and that $\frac{n}{\sigma^2} = I(\theta)$ i.e. CRLB = $\frac{\sigma^2}{n}$. Thus \bar{X} is UMVUE.

2. When μ is known but σ^2 is unknown, $\theta = \sigma^2$ and

$$\log f_X(X|\theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\theta) - \frac{1}{2\theta} \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2$$

Hence

$$\begin{aligned} S(X) &= \frac{\partial}{\partial \theta} \log f_X(X|\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta} \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2 \\ &= \frac{n}{2\theta^2} \left[\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 - \theta \right] \end{aligned}$$

$\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2$ is an unbiased estimator of $\tau = \sigma^2$ and $\frac{n}{2\theta^2} = I(\theta)$ i.e. the CRLB = $\frac{2\theta^2}{n} = \frac{2\sigma^4}{n}$

3.and 4. Case both μ and σ^2 is unknown

here $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ i.e. $\theta_1 = \mu$ and $\theta_2 = \sigma^2$

$$f_X(X|\theta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right) \propto \theta_2^{-n/2} \exp\left(\frac{1}{2\theta_2} \sum_{i=1}^n (x_i - \theta_1)^2/\sigma^2\right)$$

and

$$\log f_X(X|\theta) = -\frac{n}{2} \log \theta_2 - \frac{1}{2\theta_2} \sum_{i=1}^n (x_i - \theta_1)^2/\sigma^2$$

Thus

$$= \frac{\partial}{\partial \theta} \log f_X(X|\theta) = \frac{1}{\theta_2} \sum_{i=1}^n (x_i - \theta_1)/\sigma^2$$

$$\frac{\partial^2}{\partial \theta^2} \log f_X(X|\theta) = -\frac{n}{\theta_2}$$

$$\frac{\partial^2}{\partial \theta^2 \theta^1} \log f_X(X|\theta) = -\frac{1}{\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)$$

$$\frac{\partial}{\partial \theta^2} \log f_X(X|\theta) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)^2$$

$$\frac{\partial^2}{\partial \theta^2} \log f_X(X|\theta) = \frac{n}{2\theta_2^2} - \frac{1}{\theta_2^3} \sum_{i=1}^n (x_i - \theta_1)^2$$

Consequently

$$I_{11}(\theta) = -E\left(-\frac{n}{\theta_2}\right) = \frac{n}{\theta_2}$$

$$I_{12}(\theta) = -E\left(-\frac{1}{\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)\right) = 0$$

$$I_{22}(\theta) = -E\left(\frac{n}{2\theta_2} - \frac{1}{\theta_2^3} \sum_{i=1}^n (x_i - \theta_1)^2\right) = \frac{n}{2\theta_2^2}$$

i.e

$$I(\theta) = \begin{bmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{bmatrix}$$

and

$$[I(\theta)]^{-1} = J(\theta) = \begin{bmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Consequently, for unbiased estimators $\hat{\mu}$, $\hat{\sigma}^2$ of μ and σ^2 respectively

$$\text{Var}(\hat{\mu}) \geq \frac{\sigma^2}{n}$$

and

$$\text{Var}(\hat{\sigma}^2) \geq \frac{2\sigma^4}{n}$$

6.6.2.2 The vector case

■ For the normal example, we consider $\theta = (\mu, \sigma^2)$, where the unknown parameter is a vector.

■ Let $\theta = (\theta_1, \dots, \theta_k)$, then the score function is

$$S_n(\theta) = \left(\frac{\partial}{\partial \theta_1 l(\theta)}, \frac{\partial}{\partial \theta_2 l(\theta)}, \dots, \frac{\partial}{\partial \theta_k l(\theta)} \right)^T$$

$E[S_n(\theta)] = 0$ still holds.

■ The Fisher information is now a $k \times k$ matrix, actually, the covariance matrix for $S_n(\theta)$, that

$$I_n = E_\theta[S_n(\theta)(S_n(\theta))^T],$$

For the (r, s) element of I_n , there is $I_n(r, s) = -E_\theta \left[\frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right]$. So, under regular conditions, I_n equals to the expectation of the Hessian matrix for $-l(\theta)$.

- **Example 7.3.14 (Normal Variance Bound)** Let X_1, \dots, X_n be i.i.d. $N(\mu, \sigma^2)$. The conditions of Cramér-Rao Theorem are satisfied. Let W be an unbiased estimator of σ^2 , then

$$\text{Var}(W|\mu, \sigma^2) \geq 2\sigma^4/n.$$

In Example 7.3.3 we see that $\text{Var}(S^2|\mu, \sigma^2) \geq \frac{2\sigma^4}{n-1}$. So S^2 does not attain the Cramér-Rao lower bound.

- **Corollary 7.3.15 (Attainment)** Let X_1, \dots, X_n be i.i.d. $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramér-Rao Theorem. Let $L(\theta|\mathbf{x})$ denote the likelihood function. If $W(\mathbf{X})$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramér-Rao lower bound if and only if

$$a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial \theta} \log L(\mathbf{x}|\theta)$$

for some function $a(\theta)$.

- **Proof** The Cramér-Rao inequality, can be written as

$$\begin{aligned} & \left[\text{Cov}_\theta \left(W(\mathbf{X}), \frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i|\theta) \right) \right]^2 \\ & \leq \text{Var}_\theta W(\mathbf{X}) \cdot \text{Var}_\theta \left(\frac{\partial}{\partial \theta} \log L(\mathbf{X}) \right) \end{aligned}$$

Using the condition for “=” in Cauchy-Schwarz inequality, we obtain the expression (6.1).

• **Example 7.3.16 (Continuation of Example 7.3.14)**

$$L(\mathbf{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2\right)$$

and hence

$$\frac{\partial}{\partial \sigma^2} \log L(\mathbf{x}|\mu, \sigma^2) = \frac{n}{2\sigma^4} \left(\sum_{i=1}^n \frac{(x_i - \mu)^2}{n} - \sigma^2 \right)$$

Taking $a(\sigma^2) = \frac{n}{2\sigma^4}$ shows that the best unbiased estimator of σ^2 is $\sum_{i=1}^n (x_i - \mu)^2 / n$, which is calculable only if μ is known.

- So the question of finding best unbiased estimator are still unsolved for many common pdf's.

6.6.3 Mean Square Error

The **mean square error (MSE)** of an estimator W of a parameter θ is the function of θ defined by

$$E_{\theta}(W - \theta)^2.$$

$\text{Bias}_\theta W = E_\theta W - \theta$. If $\text{Bias}_\theta W = 0$, then W is unbiased.

Example 7.3.3 (Normal MSE) Let X_1, X_2, \dots, X_n be i.i.d. $N(\mu, \sigma^2)$. Then statistics \bar{X} and S^2 are both unbiased.

$$\begin{aligned} \text{MSE}(\bar{X}) &= E(\bar{X} - \mu)^2 = \text{Var}(\bar{X}) = \sigma^2/n \\ E(S^2 - \sigma^2)^2 &= \text{Var}(S^2) = \frac{2\sigma^4}{n-1} \end{aligned}$$

- **Example 7.3.4** Maximum Likelihood estimator of σ^2 is $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$.

$$\begin{aligned} \text{Var}\left(\frac{n-1}{n} S^2\right) &= \frac{(n-1)^2}{n^2} \cdot \frac{2\sigma^4}{n-1} = \frac{2(n-1)}{n^2} \sigma^4 \\ \text{MSE}\left(\frac{n-1}{n} S^2\right) &= \left(\frac{n-1}{n} E S^2 - \sigma^2\right)^2 + \frac{2(n-1)}{n^2} \sigma^4 \\ &= \sigma^4 \left(\frac{n-1}{n} - 1\right)^2 + \frac{2(n-1)}{n^2} \sigma^4 \\ &= \sigma^4 \frac{2n-1}{n^2} \end{aligned}$$

Since

$$\frac{2n-1}{n^2} < \frac{2}{n-1},$$

So in this case MLE has smaller MSE than the unbiased estimator S^2 .

Remark While MSE is a reasonable measurement for location parameters, it may not be a good to compare estimators of scale parameters with MSE.

Example 7.3.5 (MSE of binomial Bayes Estimator) $X_1, \dots, X_n \sim \text{Bernoulli}(p)$.

- Let $\hat{p} = \frac{X_1 + \dots + X_n}{n}$. $E_p(\hat{p} - p)^2 = \text{Var}_p(\bar{X}) = \frac{p(1-p)}{n}$.
- Let $\hat{p}_B = \frac{Y + \alpha}{\alpha + \beta + n}$ be the Bayes estimator. Here $Y = \sum_{i=1}^n X_i$

$$\begin{aligned} \text{MSE}(\hat{p}) &= \text{Var}_p(\hat{p}_B) + (\text{Bias}_p(\hat{p}_B))^2 \\ &= \text{Var}\left(\frac{Y + \alpha}{\alpha + \beta + n}\right) + \left(E_p\left(\frac{Y + \alpha}{\alpha + \beta + n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha + \beta + n)^2} + \left(\frac{np + \alpha}{\alpha + \beta + n} - p\right)^2 \end{aligned}$$

In the absence of good prior information about p , we might choose α and β to make the MSE

of \hat{p}_B constant. Choose $\alpha = \beta = \sqrt{n/4}$ gives

$$\hat{p}_B = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}, \quad E(\hat{p}_B - p)^2 = \frac{n}{4(n + \sqrt{n})^2}$$

Figure 7.3.1 Comparison of $MSE(\hat{p})$ and $MSE(\hat{p}_B)$ for sample size $n = 4$ and $n = 400$ in Example 7.3.5

- For small n , \hat{p}_B is the better choice (unless there is a strong belief that p is near 0 or 1)
- For large n , \hat{p} is the better choice (unless there is a strong belief that p is close to $\frac{1}{2}$)

第 6 讲 练习

1. Consider a random sample $X_1, X_2, \dots, X_n \sim Unif(0, 1)$.
 - (a). Find the estimator for θ through MoM, denoted by $\hat{\theta}_{MM}$.
 - (b). Find the MLE $\hat{\theta}_{MLE}$.
 - (c). What are the expectations and variances of $\hat{\theta}_{MM}$ and $\hat{\theta}_{MLE}$? Which estimator is better?

第7讲 Asymptotic properties of estimators

内容提要

- Consistency
- Efficiency and Relative Efficiency
- MLE
- Robustness

7.1 Consistency

- Recall: with LLN, the sample average satisfies that $\bar{X}_n \xrightarrow{P} E[X]$
- Generally, for an estimator $\hat{\theta}$, we hope $\hat{\theta} \xrightarrow{P} \theta$

定义 7.1 (Definition 10.1.1: Consistency)

Let $\hat{\theta}_n$ be an estimator for θ . The estimator is said to be consistent if

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



To show consistency, we can:

- Prove that $P(|\hat{\theta}_n - \theta| > \epsilon) \rightarrow 0$ for any $\epsilon > 0$
- Recall that L^p convergence indicate convergence in probability, so if $\hat{\theta}_n \xrightarrow{L^2} \theta$, $\hat{\theta}_n$ is consistent. The L^2 distance between $\hat{\theta}_n$ and θ is

$$\int (\hat{\theta}_n - \theta)^2 dF(\hat{\theta}) = E[(\hat{\theta}_n - \theta)^2] = \text{MSE}(\hat{\theta}_n)$$

So, if $\text{MSE}(\hat{\theta}_n) \rightarrow 0$, $\hat{\theta}_n$ is consistent.

例题 7.1 Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$.

- Consider the MLE $\hat{p} = \bar{X}_n$. According to LLN, $\bar{X}_n \xrightarrow{P} p$. So, the MLE \hat{p} is consistent.
- Consider the function $\tau(p) = \log(p/(1-p))$. Let the estimator be $W = \log(\bar{X}_n/(1-\bar{X}_n))$. According to the continuous mapping theorem, $W \xrightarrow{P} \tau(p)$. So W is consistent
- Consider the estimator

$$\hat{p} = \frac{\sum X_i + 1}{n + 1}.$$

For \hat{p} , we have that

$$\text{Bias}(\hat{p}) = E[\hat{p}] - p = \frac{1-p}{(n+1)} \rightarrow 0 \quad \text{Var}(\hat{p}) = \frac{np(1-p)}{(n+1)^2} \rightarrow 0.$$

Therefor, $MSE(\hat{p}) = Bias^2 + Var \rightarrow 0$. \hat{p} is consistent.

定理 7.1 (Theorem 10.1.3)

If $\hat{\theta}$ is a sequence of estimations of θ satisfying

$$(1) \lim_{n \rightarrow \infty} Var \hat{\theta} = 0$$

$$(2) \lim_{n \rightarrow \infty} E \hat{\theta} = \theta$$

then $\hat{\theta}$ is a consistent sequence of estimators of θ



- Consistency \neq Unbiasedness
 - \hat{p} is biased, but consistent
 - X_1 is unbiased for $E[X]$, but not consistent
 - In reality, we prefer consistency (when we have more samples, we can be closer to the truth)
- According to LLN, the MoM estimators are always consistent
(the random sample should satisfy the conditions for LNN)
- For Bayes model, it depends on the prior. If the prior is inappropriate, then the estimator is not consistent. (Example in tutorial)
- How about the MLE?

7.2 Consistency of MLE

定理 7.2 (Theorem 10.1.6 Consistency of MLE)

] Let X_1, X_2, \dots be i.i.d. $f(x; \theta)$, and let $L(\theta)$ be the likelihood function. Let θ be the MLE of θ . If

- θ is identifiable, i.e., if $\theta_1 \neq \theta_2$, then $f(x|\theta_1) \neq f(x|\theta_2)$; and
- $f(x; \theta)$ have common support w.r.t. different θ , and differentiable in θ ; and
- the true parameter θ_0 is an interior point of the parameter space then for any continuous function of θ , $\tau(\theta)$, there is

$$\tau(\hat{\theta}) \xrightarrow{P} \tau(\theta)$$



- If the function is $\tau(\theta) = \theta$, then there is $\hat{\theta} \xrightarrow{P} \theta$, which shows the consistency of MLE
- According to the invariance of MLE, $\tau(\hat{\theta})$ is the MLE $\tau(\theta)$. So the consistency of MLE holds even for a function of θ

证明

To prove it, we define a new term as the expectation of the one sample log-likelihood function:

$$l(\theta) = E[\log f(X_i; \theta)].$$

The sketch of the proof is

- The MLE $\hat{\theta}$ is the maximiser of the log-likelihood function $l_n(\theta)$, and also $\frac{1}{n}l_n(\theta)$
- The true parameter θ_0 is the maximiser of $l(\theta)$
- For any θ , $\frac{1}{n}l_n(\theta) \xrightarrow{P} l(\theta)$

Combine these conclusions, with the regularity conditions above and the technique in mathematical analysis, we can have the result.

In this class, we show these 3 conclusions only. The last part is easier for compact parameter space, but quite hard when we set such flexible conditions.

- The MLE $\hat{\theta}$ is the maximiser of the log-likelihood function $l_n(\theta)$. \Leftarrow Definition of MLE.

- The true parameter θ_0 is the maximiser of $l(\theta)$.

$$\begin{aligned}
 l(\theta_0) - l(\theta) &= E[\log f(X; \theta_0)] - E[\log f(X; \theta)] \\
 &= \int (\log f(X; \theta_0)) f(x; \theta_0) dx - \int (\log f(X; \theta)) f(x; \theta_0) dx \\
 &= \int \log \frac{f(X; \theta_0)}{f(X; \theta)} f(x; \theta_0) dx \\
 &= K(f(X; \theta_0), f(X; \theta)) > 0.
 \end{aligned}$$

- For any θ , $\frac{1}{n}l_n(\theta) \xrightarrow{P} l(\theta)$. Note that

$$\frac{1}{n}l_n(\theta) = \frac{1}{n} \sum \log f(X_i; \theta) \xrightarrow{P} E[\log f(X; \theta)] = l(\theta)$$

The convergence comes from WLLN.

例題 7.2(Inconsistency of an MLE) Example. Let $Y_{11}, Y_{12} \sim N(\mu_1, \sigma^2), Y_{21}, Y_{22} \sim N(\mu_2, \sigma^2), \dots, Y_{n1}, Y_{n2} \sim N(\mu_n, \sigma^2)$ Then in this case, the number of parameters increase as n increases, different from the case we discussed before, that the parameter space is fixed.

Solution. The MLE for σ^2 in this problem is

$$\hat{\sigma}^2 = \sum_{i=1}^n \sum_{j=1}^2 \frac{(Y_{ij} - \bar{Y}_i)^2}{2n}, \quad \bar{Y}_i = (Y_{i1} + Y_{i2})/2$$

which is the average of the MLE of σ^2 in each small group.

Now, for each item in the summation, note that

$$\begin{aligned} \hat{\sigma}_i^2 &= \sum_{j=1}^2 (Y_{ij} - \bar{Y}_i)^2 = (Y_{i1} - \frac{Y_{i1} + Y_{i2}}{2})^2 + (Y_{i2} - \frac{Y_{i1} + Y_{i2}}{2})^2 \\ &= \frac{(Y_{i1} - Y_{i2})^2}{2} \end{aligned}$$

So, with LLN, $\hat{\sigma}^2 = \frac{1}{2n} \sum \hat{\sigma}_i^2 = \frac{1}{n} \sum \frac{(Y_{i1} - Y_{i2})^2}{2} \xrightarrow{P} \sigma^2/2$. It does not converge to σ^2 in probability!

The modified estimator $2\hat{\sigma}^2$ is consistent.

7.3 Asymptotic Properties

- Unbiasedness \longrightarrow Asymptotic unbiasedness

$$E[\hat{\theta}_n] - \theta \rightarrow 0, \quad n \rightarrow \infty.$$

- Unbiasedness is not that important in the asymptotic theory, since we have the consistency property

- Multiple consistent estimators?
- Crámer-Rao Lower Bound \longrightarrow Asymptotic efficient

$$\text{Var}(\hat{\theta}_n) \rightarrow \text{Crámer - Rao Lower Bound}, \quad n \rightarrow \infty$$

- Asymptotic relative efficiency: comparing the variance of these two estimators

7.3.1 Asymptotic Variance

定义 7.2 (Asymptotic Variance)

Let $\hat{\theta}_n$ be an estimator for $\theta = \theta_0$. If for some deterministic sequence (a_n) , we have

$$a_n(\hat{\theta}_n - \hat{\theta}_0) \xrightarrow{d} N(0, \sigma^2)$$

The σ^2 is called the asymptotic variance.



- If we study $\hat{\theta}_n - \hat{\theta}_0$ directly, then we have that it converges to 0 in prob. since it is consistent. That does not provide more information to us.
- Usually, for (a_n) , we take it as n^c , which increases w.r.t. n , without any constant term that would impact the asymptotic variance.

例题 7.3 In CLT, the average follows the estimation is that

$$\sqrt{n}(\bar{X}_n - E[X]) \xrightarrow{d} N(0, \text{Var}(X_1))$$

Here, $a_n = \sqrt{n} = n^{1/2}$ and the asymptotic variance is $\text{Var}(X_1)$. If $a_n = n^c$ where $c < 1/2$, then $\text{Var}(a_n \bar{X}_n) \rightarrow 0$. If $a_n = n^c$ where $c > 1/2$, then $\text{Var}(\sqrt{n} \bar{X}_n) \rightarrow \infty$

注 The asymptotic variance is different from $\lim_{n \rightarrow \infty} \text{Var}(a_n \hat{\theta}_n)$ ([Definition 10.1.7 Limiting Variances](#))

例题 7.4 Example 10.1.10 (Large-sample mixture variances) The hierarchical model

$$\begin{aligned} Y_n | W_n = w_n &\sim N(0, w_n + (1 - w_n)\sigma_n^2) \\ W_n &\sim \text{Bernoulli}(p_n) \end{aligned}$$

where the sequence σ_n^2, p_n are known.

Now,

$$\begin{aligned} \text{Var}(Y_n) &= E[\text{Var}(Y_n | W_n)] + \text{Var}(E[Y_n | W_n]) \\ &= E(W_n + (1 - W_n)\sigma_n^2) \\ &= p_n + (1 - p_n)\sigma_n^2 \end{aligned}$$

The variance would converge **only if** $\lim_{n \rightarrow \infty} (1 - p_n)\sigma_n^2 < \infty$.

Now we consider the asymptotic variance. First we should figure out the distribution it converges to in distribution. For some fixed a ,

$$P(Y_n \leq a) = E[P(Y_n \leq a | W_n)] = (1 - p_n)\Phi(a/\sigma_n) + p_n\Phi(a),$$

where $\Phi(\cdot)$ is the CDF for standard normal distribution.

Therefore, if $p_n \rightarrow 1$ then $Y_n \rightarrow N(0, 1)$, so that the asymptotic variance is 1. However, if $\lim_{n \rightarrow \infty} (1 - p_n) \sigma_n^2 = \infty$ then this is the value of the limiting variance, which is of course different.

7.3.2 Asymptotic Efficiency

定义 7.3

The estimator $\hat{\theta}_n$ is asymptotic efficient for a parameter $\theta = \theta_0$ if

- $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, v(\theta))$ and
-

$$v(\theta_0) = \frac{1}{I(\theta_0)} \quad \text{Crámer - Rao Lower Bound}$$



- One estimator is efficient, as long as it's asymptotic variance exists, and meets the CRLB.
- Obviously, for $\tau(\theta)$, the definition also works, except that the CRLB becomes the CRLB for $\tau(\theta)$
- MLE is always asymptotic efficient

7.3.3 Asymptotic Normality of the MLE

Regularity conditions:

- the dimension of the parameter space does not change with n ;
- $f(x, \theta)$ have common support w.r.t. different θ , and is differentiable in θ ;
- the differentiation w.r.t. θ is interchangeable with integration over x .

定理 7.3 (Theorem 10.1.12: Asymptotic Normality of MLE)

Let $\hat{\theta}_n$ be the MLE for the parameter θ . Under the regularity condition, there is

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \frac{1}{I(\theta)}).$$

- Hence, $\hat{\theta}_n = \theta + O_p(1/\sqrt{n})$.



Proof. By Taylor's theorem

$$l'(\hat{\theta}) = l'(\theta) + (\hat{\theta} - \theta)l''(\theta) + \dots$$

Recall that $l'(\hat{\theta}) = 0$ since $\hat{\theta}$ is MLE, so we have

$$0 = l'(\theta) + (\hat{\theta} - \theta)l''(\theta) + \dots$$

$$(\hat{\theta} - \theta) \approx -\frac{l'(\theta)}{l''(\theta)}$$

$$\sqrt{n}(\hat{\theta} - \theta) \approx -\frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)} \equiv \frac{A}{B}$$

Now, $A = \frac{1}{\sqrt{n}}l'(\theta) = \sqrt{n} \times \frac{1}{n} \sum_{i=1}^n S(\theta, X_i) = \sqrt{n}(\bar{S}_n - 0)$, where $S(\theta, X_i)$ is the score function based on X_i . Recall that $E[S(\theta, X_i)] = 0$ and $Var(S(\theta, X_i)) = I(\theta)$. By CLT, $A \xrightarrow{d} N(0, I(\theta))$.

By WLLN, $B \xrightarrow{P} E[l''(\theta)] = I(\theta)$. Combine them, by Slutsky's theorem,

$$\sqrt{n}(\hat{\theta} - \theta) \approx \frac{A}{B} \xrightarrow{d} \frac{1}{I(\theta)} N(0, I(\theta)) = N(0, 1/I(\theta)).$$

- According to the proof, we can see that

$$\hat{\theta} = \theta + \frac{1}{n} \sum_{i=1}^n \frac{S(\theta, X_i)}{I(\theta)} + o_p(n^{-1/2}).$$

The function $\frac{S(\theta, X_i)}{I(\theta)}$ is called the [influence function](#).

- The asymptotic variance of $\hat{\theta}$ is $1/I(\theta)$
- The estimated asymptotic variance of $\hat{\theta}$ is $1/I(\hat{\theta})$

- If τ is a smooth function of θ , then

$$\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \xrightarrow{d} N(0, (\tau'(\theta))^2 / I(\theta)),$$

with asymptotic variance $(\tau'(\theta))^2 / I(\theta)$. The estimated asymptotic variance is $(\tau'(\hat{\theta}))^2 / I(\hat{\theta})$

例题 7.5 $X_1, X_2, \dots, X_n \stackrel{i.i.d}{\sim} \text{Exponential}(\theta)$

Now $f(x; \theta) = \theta e^{-\theta x}$ and $L(\theta) = \theta^n e^{-n\theta \bar{X}}$, Hence, the log-likelihood function

$$l(\theta) = -n\theta \bar{X} + n \log \theta$$

and

$$S(\theta) = \frac{n}{\theta} - n\bar{X}, \quad l''(\theta) = -n/\theta^2 < 0$$

The MLE is $\hat{\theta} = 1/\bar{X}$. The Fisher information is $I_n(\theta) = -E[-n/\theta^2] = n/\theta^2$. Therefore,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \theta^2)$$

例题 7.6 Example. $X_1, X_2, \dots, X_n \stackrel{i.i.d}{\sim} \text{Bernoulli}(p)$

We can find that the Fisher information for $n = 1$ is $I(p) = 1/(p(1 - p))$. So, for the MLE $\hat{p} = \bar{X}$,

$$\sqrt{n}(\bar{X} - p) \xrightarrow{d} N(0, p(1 - p)).$$

Now suppose we want to estimate $\tau = p/(1 - p)$. The MLE is $\hat{\tau} = \bar{X}/(1 - \bar{X})$, according to the invariance of MLE. Now

$$\frac{\partial}{\partial p} \frac{p}{1 - p} = \frac{1}{(1 - p)^2}.$$

The asymptotic distribution for $\hat{\tau}$ is

$$\sqrt{n}(\bar{X}/(1 - \bar{X}) - p/(1 - p)) \xrightarrow{d} N\left(0, \frac{1}{(1 - p)^4} \times p(1 - p)\right) = N\left(0, \frac{p}{(1 - p)^3}\right)$$

7.4 Asymptotic Relative Efficiency

定义 7.4 (Definition 10.1.16 Asymptotic Relative Efficiency)

Suppose that two estimator W_n and V_n satisfy

$$\sqrt{n}(W_n - \tau(\theta_0)) \xrightarrow{d} N(0, \sigma_W^2)$$

$$\sqrt{n}(V_n - \tau(\theta_0)) \xrightarrow{d} N(0, \sigma_V^2)$$

then the least favorable prior is defined as

$$ARE(W_n, V_n) = \frac{\sigma_V^2}{\sigma_W^2}$$



- When W_n and V_n have the same convergence rate, the ARE compares their efficiency.

例题 7.7 Example 10.1.17 Let $X_1, \dots, X_n \stackrel{i.i.d}{\sim} \text{Poisson}(\lambda)$. The MLE of λ is \bar{X} . Let $\tau = P(X_i = 0) = e^{-\lambda}$. Define $Y_i = I(X_i = 0)$. This suggests the estimator

$$W_n = \frac{1}{n} \sum Y_i$$

Another estimator is the MLE $V_n = e^{-\bar{X}}$. Compare them.

Since $E[X_1] = \lambda$, $Var(X_1) = \lambda$, so according to CLT, there is

$$\sqrt{n}(\bar{X} - \lambda) \xrightarrow{d} N(0, \lambda).$$

According to the data method, we have

$$\sqrt{n}(V_n - e^{-\lambda}) \xrightarrow{d} N(0, e^{-2\lambda}\lambda).$$

Note that $Y_i \sim \text{Bernoulli}(e^{-\lambda})$, so

$$\sqrt{n}(W_n - e^{-\lambda}) \xrightarrow{d} N(0, e^{-\lambda}(1 - e^{-\lambda})).$$

So we have

$$ARE(W_n, V_n) = \frac{e^{-2\lambda}\lambda}{e^{-\lambda}(1 - e^{-\lambda})} = \frac{\lambda}{e^{\lambda} - 1} \leq 1.$$

So, for most λ , MLE is more efficient.

7.5 Robustness

- MLE is efficient only if the model is right. It can be vary bad if the model is wrong

Example. Suppose $X_1, \dots, X_n \stackrel{i.i.d}{\sim} N(\theta, \sigma^2)$. The MLE is $\hat{\theta}_n = \bar{X}_n$.

Suppose, we have a perturbed model that $X_i \sim N(\theta, \sigma^2)$ with probability $1 - \delta$ and $X_i \sim f(x)$ with probability δ , where

$$f(x) = \frac{1}{\pi(x^2 + 1)}.$$

This is the Cauchy distribution, which is quite famous as an example that $E[X] = +\infty$. Therefore, $Var(\bar{X}) = \infty$.

If we still apply \bar{X}_n , then the Cauchy distribution will destroy its good properties. However, for small δ , the median still keeps the same. On the other hand, if we consider the normal model as correct, then in the next slide we can show that $ARE(M_n, mle) = 0.64 < 1$, which indicates that MLE is better.

- Nonparametric estimation (say, the median) is a solution
- Even when the model is wrong, sometimes the MLE still provides some information

Find the asymptotic distribution for the median of X , assuming the model is $N(\theta, 1)$. (Let $\sigma^2 = 1$ for simplicity. The result is the same for any σ^2)

For fixed a , Let $Y_i = I(X_i \leq \theta + a/\sqrt{n})$. Then $Y_i \sim \text{Bernoulli}(p_n)$, where

$$p_n = \Phi(\theta + a\sqrt{n}) = \Phi(\theta) + \frac{a}{\sqrt{n}}\phi(\theta) + o(n^{-1/2}) = \frac{1}{2} + \frac{a}{\sqrt{n}}\phi(\theta) + o(n^{-1/2})$$

Also, $\sum_i Y_i$ has mean np_n and standard $\sigma_n = \sqrt{np_n(1-p_n)}$.

Note that, $M_n \leq \theta + a/\sqrt{n}$ if and only if $\sum Y_i \geq \frac{n+1}{2}$.

Then

$$\begin{aligned} P(\sqrt{n}(M_n - \mu) \leq a) &= P(M_n \leq \theta + a/\sqrt{n}) = P(\sum Y_i \geq \frac{n+1}{2}) \\ &= P\left(\frac{\sum Y_i - np_n}{\sigma_n} \geq \frac{(n+1)/2 - np_n}{\sigma_n}\right). \end{aligned}$$

Now, $\frac{(n+1)/2 - np_n}{\sigma_n} \rightarrow -2af(\theta)$, and hence

$$P(\sqrt{n}(M_n - \mu) \leq a) \rightarrow P(N(0, 1) \geq -2af(\theta)) = P\left(\frac{N(0, 1)}{2f(\theta)} \leq a\right),$$

so that $\sqrt{n}(M_n - \theta) \xrightarrow{d} N(0, \frac{1}{4f(\theta)^2})$, and $ARE(M_n, mle) = 0.64$.

Rcode

```

1 rm(list=ls())
2 N = 50; n = 100; sigma = 3; alpha = 0.05;
```

```
3  iter = 100; fwr1 = rep(0, iter); fwr2 = rep(0, iter);
4  for(i in 1:iter){
5    Y = rnorm(N, mean = 0, sd = sigma/sqrt(n));
6    #Generate Y_i' s under null
7    stat = sqrt(n)*Y/sigma;
8    #Calculate the test statistic T_i
9    p = 1*(abs(stat) > qnorm(1 - alpha/2))
10   #Find the p-value for each individual test without correction
11   corp = 1*(abs(stat) > qnorm(1 - alpha/2/N))
12   #Find the p-value for each individual test with Bonferroni cor
13   fwr1[i] = 1*(sum(p) > 0); #Familywise error for test 1;
14   fwr2[i] = 1*(sum(corp) > 0); #Familywise error for test 2;
15 }
16 mean(fwr1) #empirical familywise error for uncorrected test
17 mean(fwr2) #empirical familywise error for corrected test
```

7.5.1 Familywise Error Control

Familywise Error Control

- To control the overall false positives, we consider

$$P(\text{making at least one false rejection})$$

- Define $I = \{i; H_{i0} \text{ is true}\}$ be the index set for which H_0 is true
- Define $R = \{i; H_{j0} \text{ is rejected}\}$ be the index set that we reject.
- Define the familywise error rate at level α if

$$P(R \cap I \neq \emptyset) = P(\text{making at least one false rejection}) \leq \alpha.$$

- **Bonferroni method:** for each individual test, set the level to be α/N . Let p_j be the p -value for test H_{j0} versus H_{j1} .

$$\begin{aligned} P(\text{making a false rejection}) &= P(p_j < \alpha/N \text{ for some } i \in I) \\ &\leq \sum_{i \in I} P(p_j < \alpha/N) \\ &= \sum_{i \in I} \alpha/N = \frac{\alpha|I|}{N} \leq \alpha \end{aligned}$$

So we have overall control of the type 1 error.

- It can have low power.

7.5.2 Familywise Error Control

False Discovery Control

- Recall we have the table:

	Decision	
	Retain H_0	Retain H_0
H_0 is true	✓ (true negative)	Type 1 error(false positive)
H_1 is true	Type 2 error(false positive)	✓ (true postive)

- Define the false discovery proportion as

$$FDP = \frac{|R \cup I|}{|R|} = \frac{\#FP}{\#FP + \#TP}$$

- The false discovery rate is defined as the expectation of FDP.

Our goal is to let

$$FDR = E[FDP] \leq \alpha$$

False Discovery Control Benjamin-Hochberg method:

- (1) Find the ordered p -values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$
- (2) Let $j = \max\{i : p_{(i)} < i\alpha/N\}$. Let $T = p_{(j)}$.
- (3) Let $R = \{i : p_i \leq T\}$.

Proof. For simplicity, assume we reject hypothesis j when $p_j \leq t$. Let \hat{G} be the empirical distribution of the p -values and let $G(t) = E[\hat{G}(t)]$. In this case,

$$FDP = \frac{\sum_{i=1}^N I(p_i < t, H_0)}{\sum_{i=1}^N I(p_i < t)} = \frac{\frac{1}{N} \sum_{i=1}^N I(p_i < t, H_0)}{\frac{1}{N} \sum_{i=1}^N I(p_i < t)}.$$

Hence,

$$\begin{aligned} E[FDP] &\approx \frac{E[\frac{1}{N} \sum_{i=1}^N I(p_i < t, H_0)]}{\frac{1}{N} E[\sum_{i=1}^N I(p_i < t)]} = \frac{\frac{1}{N} \sum_{i=1}^N E[I(p_i < t, H_0)]}{\frac{1}{N} \sum_{i=1}^N E[I(p_i < t)]} \\ &= \frac{t\#P/N}{G(t)} \leq \frac{t}{G(t)} \approx \frac{t}{\hat{G}(t)} \end{aligned}$$

Let $t = p_{(i)}$ for some i ; then $\hat{G}(t) = i/N$. Thus, $FDR \leq p_{(i)}N/i$. Setting it equal to α and we have the result.

7.5.3 Higher Criticism

Higher Criticism

- Let p_j be the p -value for test problem H_j . Then under null hypothesis, $p_j \sim \text{Unif}(0, 1)$.
- Consider a level α test for an individual hypothesis test, we reject the hypothesis when $p_j < \alpha$
- Let $Y_j = I(p_j \leq \alpha)$. Under null hypothesis, $Y_j \sim \text{Bernoulli}(\alpha)$. So \bar{Y}_n has mean α and standard deviation $\sqrt{\alpha(1-\alpha)/N}$
- According to CLT, let \tilde{Y}_N be the fraction of the rejected hypothesis with a level α test, then

$$T_N = \sqrt{N} \frac{\tilde{Y}_n - \alpha}{\sqrt{\alpha(1-\alpha)}} \xrightarrow{d} N(0, 1),$$

When $|T_n| > z_{\alpha/2}$, at least one hypothesis testing problem rejects H_0

Higher Criticism

- Why do we consider a fixed α ? When α changes, will we get different results?
- Define the function of α . The expression is as following:
 - (1) Sort p -values so that $p_{(1)} < p_{(2)} < \cdots < p_{(N)}$

(2) Define

$$HC_k = \sqrt{N} \frac{k/N - p_{(k)}}{\sqrt{p_{(k)}(1 - p_{(k)})}}.$$

(3) The test statistic is $T_N = \max_{1 \leq k \leq N/2} HC_k$

- Note. Now $p_{(k)}$ plays the role of α . If we take $\alpha = p_{(k)}$, then the number of rejected hypothesis is k , so the fraction is k/N .
- The limiting distribution for T_n is Gumbel distribution (no need to know)

Comparison

- The three methods cares about different things
- The Bonferoni correction is to control the familywise error, which is the exact number of Type 1 wrong decisions. It does not need independence assumption between tests, but it may lose some power.
- The FDR method cares about the fraction of false positive. It is useful when the true positives are rare.
- Higher Criticism cares about whether there is any positive or not. It works for the case that true positives are rare, and it allows the signals to be moderately weak. On the other hand, to

make sure it works, the dependence between tests cannot be strong.

第 8 讲 Hypothesis Testing

- We do not need good estimation of the parameter; we are interested in one value only
- To play a game, we are wondering whether the die is *fair* or not. We do not care how *unfair* it is
- To test the effects of two medicine, we are interested in the difference of the effect equals to 0 or not

Formalize it and we state it as a null hypothesis H_0 and an alternative hypothesis H_1 . For example,

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

Generally, we want to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

Where $\Theta_0 \cap \Theta_1 = \emptyset$. If $\Theta_0 = \{\theta\}$, it is called a simple null hypothesis, otherwise, it is a complex null hypothesis

For a hypothesis testing problem:

- Underlying truth: H_0 is true or H_1 is true

- Goal: sufficient evidence to reject H_0 ?
- Action: reject H_0 or not reject H_0

	Decision	
	Retain H_0	Retain H_0
H_0 is true	✓	Type 1 error(false positive)
H_1 is true	Type 2 error(false positive)	✓

- Without sufficient evidence, we do not reject H_0 . It does not mean we believe it is correct
- Obviously, the setting prefers H_0

例题 8.1 $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. Then the problem is

$$H_0 : p = 1/2 \quad \text{versus} \quad H_1 : p \neq 1/2.$$

- What is a test?
- A test need a statistic T and a rejection region R . If $T \in R$ then we reject H_1 .

- For example, let $T = \bar{X}$ and the rejection be $(0, 0.3) \cup (0.6, 1)$, then the test is

$$\text{Reject } H_0 \text{ if } |\bar{X} - 1/2| > 0.1.$$

- With the data, we can claim whether we reject H_0 or not
- With this test, Type 1 error is $P(|\bar{X} - 1/2| > 0.1 | H_0)$, Type 2 error is $P(|\bar{X} - 1/2| > 0.1 | H_1)$
- There are multiple tests for one hypothesis testing problem

8.1 Evaluation of a test

- With a test, we hope we can do correct justifications.
- It means minimizing the Type 1 error and Type 2 error.
- For the type 2 error, we define the power function.

定义 8.1 (Power function)

] Suppose we reject H_0 when $T(X_1, \dots, X_n) \in R$. The power function is defined as

$$\beta(\theta) = P(T(X_1, \dots, X_n) \in R | \theta).$$



Remark.

- The power function is a function about θ
- When $\theta \in \Theta_1$, it measures the probability that the test correctly rejects H_0
- When $\theta \in \Theta_0$, it measures the type 1 error.
- For the type 1 error, one way is to control it with the maximum

定义 8.2 (Definition 8.3.5 and 8.3.6)

A test is size α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$$

A test is level α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$$



- A *size α* test and a *level α* test are almost the same thing. The distinction is made because sometimes we want a *size α* test and we cannot construct a test with **exact** *size α* . But we can build one with smaller error rate
- Motivation: Type 1 error is not the same important with the Type 2 error. say, for medical

diagnosis, we should minimize the Type 2 error(discover people with disease correctly), and control Type 1 error (healthy people are labeled with disease) at a low level.

- Common values for α : 0.01, 0.05, 0.1

8.2 Evaluation

The general strategy to construct a test is

- (1) Fixe $\alpha \in [0, 1]$
- (2) Try to maimize $\beta(\theta)$ for $\theta \in \Theta_0$, subject to $\beta(\theta) \leq \alpha$ for $\theta \in \Theta_0$

例题 8.2 $X_1, \dots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Suppose we test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

This is called a one – sided alternative. Suppose we reject H_0 if $T_n > c$ where

$$T_n = \frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}}.$$

Then, the power function is

$$\begin{aligned}\beta(\theta) &= P\left(\frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} > c; \theta\right) = P\left(\frac{\bar{X}_n - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}; \theta\right) \\ &= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right),\end{aligned}$$

where $Z \sim N(0, 1)$ and Φ is the CDF for Z . Now,

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = 1 - \Phi(c).$$

To get a *size* α test, set $1 - \Phi(c) = \alpha$ so that

$$c = z_\alpha = \Phi^{-1}(1 - \alpha).$$

Our test is to reject H_0 when $T_n = \frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} > z_\alpha$.

例題 8.3 Now, let's consider the **two-sided alternative**, that

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

We will reject H_0 if $|T_n| > c$. The power function is

$$\begin{aligned}
 \beta(\theta) &= P(T_n < -c; \theta) + P(T_n > c; \theta) \\
 &= P\left(\frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} < -c; \theta\right) + P\left(\frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} > c; \theta\right) \\
 &= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
 &= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + \Phi\left(-c - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)
 \end{aligned}$$

The size is $\beta(\theta_0) = 2\Phi(-c)$. Let it equal to α , then

$c = -\Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}$. The test is to reject H_0 when $|\frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}}| > z_{\alpha/2}$

8.3 Generally used tests

There are some tests that are found to be useful or optimal:

- The Neyman-Pearson Test
- The Wald Test
- The Likelihood Ratio Test (LRT)

- The permutation test

Now we discuss them one by one.

8.3.1 The Neyman-Pearson Test

- The Neyman-Pearson test considers only simple null and simple alternative, which means the test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

Definition: Neyman-Pearson Test

Let $L(\theta) = f(X_1, \dots, X_n; \theta)$ and

$$T_n = L(\theta_1)/L(\theta_0).$$

Suppose we reject H_0 if $T_n > k$ where k is chosen so that

$$P(T(X_1, \dots, X_n) > k; \theta > \theta_0) = \alpha,$$

then it is called a Neyman-Pearson Test.

- The test statistic is the ratio of two joint densities. It is to check with which likelihood, the data is more possible.
- It is quite limited, since it requires both the null and the alternative are simple.

定义 8.3 (Definition 8.3.12: Uniformly Powerful Tests)

] Let C_α be a collection of level α for $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$. A test in C_α with power function $\beta(\theta)$ is uniformly most powerful (UMP) if for every $\beta'(\theta)$ which is the power function of any other test in C_α , then

$$\beta(\theta) \geq \beta'(\theta), \quad \theta \in \Theta_1.$$



- It is possible that a UMP does not exist
- In the simple null and simple alternative case, it exists, which is the Neyman-Pearson test.

Neyman-Pearson Lemma

Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Then

- The Neyman-Pearson test is a UMP level α test;
- If such a test exists, then every UMP level α test is a Neyman-Pearson test.

例題 8.4 **Example 8.3.14** Let $X_1 \sim \text{Bernomial}(2, \theta)$ and we want to test

$$H_0 : \theta = 1/2 \quad \text{versus} \quad H_1 : \theta = 3/4.$$

We have that

$$\frac{f(0; 3/4)}{f(0; 1/2)} = \frac{1}{4}, \quad \frac{f(1; 3/4)}{f(1; 1/2)} = \frac{3}{4}, \quad \frac{f(2; 3/4)}{f(2; 1/2)} = \frac{9}{4}$$

If we construct a test that reject H_0 when

$$\frac{f(X_1; 3/4)}{f(X_1; 1/2)} > 2$$

Then the test has level as

$$P\left(\frac{f(X_1; 3/4)}{f(X_1; 1/2)} > 2; \theta_0\right) = P(X_1 = 2; \theta_0) = 1/4.$$

So it is a UMP level $1/4$ test.

- The Neyman-Pearson test considers only simple null and simple alternative, which means the test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

- Test statistic:

$$T_n = L(\theta_1)/L(\theta_0).$$

Rejection region: $T_n > k$, where k is decided according to the level/size of the test

- It is uniformly most powerful (UMP) test, which means that

$$\beta(\theta) \geq \beta'(\theta), \quad \theta \in \Theta_1,$$

where $\beta'(\theta)$ is any other test with the same level/size.

8.4 The Wald Test

- Assume there is an asymptotic normal estimator $\hat{\theta}_n$, where $\hat{\theta}_n - \theta \xrightarrow{d} N(0, \sigma_n^2)$
- If $H_0 : \theta = \theta_0$ is true, then there is $\hat{\theta}_n - \theta \xrightarrow{d} N(0, \sigma_n^2)$

- we can construct a test statistic

$$T_n = \frac{\hat{\theta}_n - \theta_0}{\hat{\theta}_n}$$

- If H_0 is true, $T_n \xrightarrow{d} N(0, 1)$, which concentrates at 0. So, if T_n is too large/small, we reject H_0 .
- This kind of test is called the Wald Test.

例題 8.5 With Bernoulli data, to test $H_0 : p = p_0$ and $H_1 : p \neq p_0$, recall that $\sqrt{n}(\bar{X} - p) \xrightarrow{d} N(0, p(1 - p))$, we can construct a Wald test

$$T_n = \left| \frac{\bar{X} - p_0}{\sqrt{\hat{p}(1 - \hat{p})/n}} \right| > c,$$

where $c = \Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}$.

- Consider MLE $\hat{\theta}_n$. According to the asymptotic normality of MLE, there is

$$\sqrt{n} \frac{\hat{\theta}_n - \theta}{\sqrt{1/I(\theta)}} \xrightarrow{d} N(0, 1).$$

So we can construct a test w.r.t. MLE, which is to reject null hypothesis when

$$T_n = \frac{\hat{\theta}_n - \theta_0}{\sqrt{1/nI(\hat{\theta})}} > c.$$

- If it happens that \bar{X} is an estimator for θ . According to CLT, $\sqrt{n}(\bar{X} - \theta)/\sigma \xrightarrow{d} N(0, 1)$. So we can also build a Wald test based on the average
- Usually, σ_n is a function of θ . Since the truth is unknown, we can either apply θ_0 or $\hat{\theta}$ in practice.
- The Wald test requires asymptotic normality, so it works for *large* sample size only.

8.5 The Likelihood Ratio Test (LRT)

- Neymann-Pearson test is the ratio of likelihoods w.r.t. two values
- For composite null and alternative, we can generalize the idea

定义 8.4 (Definition 8.2.1: Likelihood Ratio Test (LRT))

] The LRT statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ is

$$\lambda(x_1, \dots, x_n) = \frac{\sup_{\theta \in \Theta_0} (f_{X_{1:n}}(x_1, \dots, x_n; \theta))}{\sup_{\theta \in \Theta} (f_{X_{1:n}}(x_1, \dots, x_n; \theta))}$$

A LRT is any test that has a rejection region of the form

$\{(x_1, \dots, x_n); \lambda(x_1, \dots, x_n) \leq c\}$ for any constant $c \in [0, 1]$.



- Θ_0 : null parameter space; Θ : the whole parameter space
- According to the definition of MLE, the LRT statistic can be written as

$$\lambda(x_1, \dots, x_n) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

- If it is small, then we reject H_0 .

例題 8.6**Example 8.2.2** Example. Suppose that $X_i \stackrel{i.i.d}{\sim} N(\theta, 1)$ and suppose we want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Recall that the MLE is $\hat{\theta} = \bar{X}_n$. So the LRT statistic is

$$\lambda(x_1, \dots, x_n) = \frac{\frac{1}{(\sqrt{2\pi})^n} e^{-\frac{\sum (x_i - \theta_0)^2}{2}}}{\frac{1}{(\sqrt{2\pi})^n} e^{-\frac{\sum (x_i - \bar{x})^2}{2}}} = \frac{\exp\{-\frac{\sum (x_i - \theta_0)^2}{2}\}}{\exp\{-\frac{\sum (x_i - \bar{x})^2}{2}\}}.$$

Since $\sum (x_i - \theta_0)^2 = \sum (x_i - \bar{x})^2 + n \sum (\bar{x} - \theta_0)^2$, we have $\lambda(x_1, \dots, x_n) = \exp\{-\frac{n}{2}(\bar{x} - \theta)^2\}$.

Since it is monotone with $|\bar{x} - \theta_0|$, so the rejection region is equivalent with

$$\{x \in \mathbb{R}^n : |\bar{x} - \theta_0| \geq c\}.$$

Since $\bar{x} - \theta_0 \sim N(0, 1/n)$ under null hypothesis, the level of the test is $2\Phi(-\sqrt{nc})$. For a level α test, we have $c = \Phi^{-1}(1 - \alpha/2)/\sqrt{n}$.

- Can we always find a proper distribution for the LRT statistic?
- Not always, but asymptotically, yes.

Theorem: LRT statistics

Let $X_i \stackrel{i.i.d}{\sim} F_X(\cdot; \theta^*)$ with $f_X(\cdot; \theta^*)$ as the associated PDF. Let $\hat{\theta}_n$ be the MLE. Consider the testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ where $\theta \in \mathbb{R}$.

The under H_0 ,

$$-2 \log(\lambda(X_1, X_2, \dots, X_n)) \xrightarrow{d} \chi^2_1.$$

- Regularity conditions for MLE normality
- To construct a level α test, we can set the rejection region as $-2 \log(\lambda(X_1, X_2, \dots, X_n)) \geq \chi^2_{1, \alpha}$, where $\chi^2_{1, \alpha}$ is the $1 - \alpha$ quantile for $\chi^2_{1, \alpha}$ distribution.
- If $\theta = (\theta_1, \theta_2, \dots, \theta_k)$. Then, under the regularity conditions,

$$T_n = -2 \log(\lambda(X_1, X_2, \dots, X_n)) \xrightarrow{d} \chi^2_v, \quad v = \dim(\Theta) - \dim(\Theta_0).$$

Proof

- Under the regularity conditions, we have the Taylor expansion for the log-likelihood function $l(\theta)$ close to the point $\hat{\theta}$:

$$l(\theta) \approx l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + l''(\hat{\theta})\frac{(\theta - \hat{\theta})^2}{2} = l(\hat{\theta}) + l''(\hat{\theta})\frac{(\theta - \hat{\theta})^2}{2}$$

- The expression for LRT statistic is

$$\begin{aligned} -2 \log(\lambda(X_1, X_2, \dots, X_n)) &= -2l(X_1, X_2, \dots, X_n; \theta_0) \\ &\quad + 2l(X_1, X_2, \dots, X_n; \hat{\theta}) \\ &\approx 2l(\hat{\theta}) - 2l(\hat{\theta}) - l''(\hat{\theta})(\theta - \hat{\theta})^2 \\ &= -l''(\hat{\theta})(\theta - \hat{\theta})^2 \\ &= \frac{-l''(\hat{\theta})}{I_n(\theta_0)} \times I_n(\theta_0)(\hat{\theta} - \theta_0)^2 = A_n \times B_n \end{aligned}$$

- Note that $A_n \xrightarrow{P} 1$ according to WLLN and $\sqrt{B_n} \xrightarrow{d} N(0, 1)$, so that $B_n \xrightarrow{d} \chi_1^2$. According to Slutsky's theorem, the result follows.

例题 8.7

- $X_1, \dots, X_n \stackrel{i.i.d}{\sim} \text{Poisson}(\lambda)$. The log-likelihood function is $\sum X_i \log \lambda - n\lambda + C$. We want to test $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$.

Recall the MLE is $\hat{\lambda} = \sum X_i/n$, then the LRT statistic is

$$-2 \log(\lambda(X_1, X_2, \dots, X_n)) = 2n[(\lambda_0 - \hat{\lambda}) - \hat{\lambda} \log(\lambda_0/\hat{\lambda})]$$

we reject H_0 when $-2 \log(\lambda(X_1, X_2, \dots, X_n)) > \chi^2_{1, \alpha}$.

例题 8.8LRT: examples Consider a multinomial distribution with $\theta = (p_1, p_2, \dots, p_5)$, So

$$L(\theta) = p^{y_1} \dots p^{y_5}, \quad y_k = \sum 1\{X_i = k\}, k = 1, 2, 3, 4, 5.$$

Suppose we want to test

$$H_0 : p_1 = p_2 = p_3 \text{ and } p_4 = p_5 \text{ versus } H_1 : H_0 \text{ is false.}$$

Then $v = \dim(\Theta) - \dim(\Theta_0) = 4 - 1 = 3$. The LRT test statistic is

$$\lambda(x_1, \dots, x_n) = \frac{\prod_{i=1}^5 \hat{p}_{0j}^{Y_j}}{\prod_{i=1}^5 \hat{p}_j^{Y_j}}$$

where $\hat{p}_j = Y_j/n$, $\hat{p}_{10} = \hat{p}_{20} = \hat{p}_{30} = (Y_1 + Y_2 + Y_3)/n$,

$\hat{p}_{40} = \hat{p}_{50} = (1 - 3\hat{p}_{10})/2$. We reject H_0 if the test statistic is larger than $\chi^2_{3, \alpha}$

8.6 p -values

- Given α , we construct a level α test
- With data, we calculate the statistic and decide whether to *reject* or *retain* H_0
- If α changes, should we do all the steps again?

定义 8.5 (Definition 8.3.26: P-values)

A p -value $p(X)$ is a test statistic with $p(X) \in [0, 1]$. Small values of p indicate that H_1 is true. A p -value is **valid** if for every $\theta \in \Theta_0, \alpha \in [0, 1]$,

$$P(p(X) \leq \alpha; \theta) \leq \alpha.$$



- $p(X)$ is a test statistic. With the statistic $p(X)$, the level α test is to reject H_0 when $p < \alpha$. The power function w.r.t. this test is

$$\beta(\theta) = P(p(X) \leq \alpha; \theta).$$

- Therefore, it can be **viewed as the smallest α at which we would reject H_0** .
- Question: how to find this test statistic?

定理 8.1 (Theorem 8.3.27: P-values)

Let $W(X)$ be a test statistic such that large values of W indicate that H_1 is true. For each $x \in X$, define

$$p(x) = \sup_{\theta \in \Theta_0} P(W(X) \geq W(x); \theta),$$

then $p(X)$ is a valid p -value.



- Note that $W(x)$ should satisfy that reject H_0 when $T(x) > c$.
- This is the general way to find the p -value. We define a test statistic first, and then define p be the probability that the statistic is no smaller than the observation.
- p -value may change for different test statistic, even with the same data. So when we specify p -value, we should specify the test statistic.
- p -value is the probability for the test statistic under null, not the probability of H_0
- Why p -value is useful, not the test statistic?

Theorem

Under $H_0, p \sim \text{Unif}(0, 1)$

We never know what the test statistic means, but we can achieve the information in p -value quickly

例题 8.9 Example. Let $X_1, \dots, X_n \stackrel{i.i.d}{\sim} N(0, 1)$. Test that $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We reject when $|T_n| = |\sqrt{n}(\bar{X}_n - \theta_0)|$ is large. Let t_n be the observed value of T_n . Let $Z \sim N(0, 1)$. Then,

$$p = P(|\sqrt{n}(\bar{X}_n - \theta_0)| > t_n) = P(|Z| > t_n) = 2\Phi(-|t_n|).$$

Now, we can return the p -value to the researcher, with which the researcher can easily tell how strong the evidence is to reject H_0 .

8.7 The Permutation Test

Suppose we have $X_1, \dots, X_n \sim F$ and $Y_1, \dots, Y_m \sim G$. We want to test

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G$$

- (1) Let $Z = (X_1, \dots, X_n, Y_1, \dots, Y_m)$. Create labels as $L = (1, 1, \dots, 1, 2, \dots, 2)$, where there are n 1's and m 2's. So L is the label for the observation.
- (2) The test statistic can be written as a function of Z and L . For example, $|\bar{X}_n - \bar{Y}_m|$ can be written as

$$T = \left| \frac{\sum_{i=1}^{m+n} Z_i I(L_i = 1)}{\sum_{i=1}^{m+n} I(L_i = 1)} - \frac{\sum_{i=1}^{m+n} Z_i I(L_i = 2)}{\sum_{i=1}^{m+n} I(L_i = 2)} \right|$$

So $T = g(L, Z)$.

- (3) Define:

$$p = \frac{1}{(n+m)!} \sum_{\pi} I(g(L_{\pi}, Z) > g(L, Z)),$$

where L_{π} is a permutation of the labels and the sum is over all permutations.

- (4) Under H_0 , $F = G$, so the distribution of T does not change, and $p \sim \text{Unif}(0, 1)$ (discrete version).
- (5) Reject H_0 if $p < \alpha$.

Summing over all permutations is infeasible for large data sets. The computation load is $(n+m)!$.

Usually, it suffices to use a random sample of the permutations. So the procedure becomes

- (1) Let $Z = (X_1, \dots, X_n, Y_1, \dots, Y_m)$. Create labels as $L = (1, 1, \dots, 1, 1, \dots, 2)$, where there

are n 1's and m 2's.

(2) Let $T = g(L, Z)$. Compute a random permutation of the labels $\pi_i, i = 1, \dots, K$. Define

$$p = \frac{1}{K} \sum_{i=1}^K I(g(L_{\pi_i}, Z) > g(L, Z))$$

(3) Reject H_0 if $p < \alpha$.

- Distribution free
- Does not involve any asymptotic approximation
- Flexible to derive for any statistics

8.8 Multiple Testing

- The classical hypothesis testing problem is to consider limited parameters, say $\theta = \theta_0$ or no
- Sometimes, we need to do multiple testing at the same time, say

$$H_{10} : \theta_1 = 0 \quad \text{versus} \quad H_{11} : \theta_1 \neq 0$$

$$H_{20} : \theta_2 = 0 \quad \text{versus} \quad H_{21} : \theta_2 \neq 0$$

... ..

$$H_{N0} : \theta_N = 0 \quad \text{versus} \quad H_{N1} : \theta_N \neq 0$$

- For example, we want to identify the genes that cause one specific disease. For each gene, we want to know whether it works or not. The number of genes is pretty large here.
- Can we do the hypothesis testing problems individually, and reject H_0 if any individual H_0 is rejected?
- No and Yes.
- Recall. For a level α test, we have $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.
- Consider the individual testing problem:

$$H_{i0} : \theta_i = 0 \quad \text{versus} \quad H_{i1} : \theta_i \neq 0$$

Suppose we have a level α test for this problem. Denote the power function of this test as $\beta_i(\theta)$.

Then $\beta_i(\theta) \leq \alpha$.

- Therefore, for the original multiple-testing problem, the rejection probability is

$$\beta(\theta) = 1 - P(\text{accept } H_{i0} \text{ for all } i \leq i \leq N) \stackrel{\text{indep}}{=} 1 - \prod_{i=1}^N (1 - \beta_i(\theta))$$

- Under null hypothesis $\theta = 0$, if all the test statistic are independent and all the tests have size α . The rejection probability is

$$1 - \prod_{i=1}^N (1 - \beta_i(0)) = 1 - (1 - \alpha)^N$$

How large it is? Let $N = 50$, $\alpha = 0.05$, then $1 - (1 - \alpha)^N \approx 0.92$.

$P(\text{rejection}|H_0) = 0.92!$

8.8.1 Familywise Error Control

We need to adjust the level for individual tests

- Define $I = \{i; H_{i0} \text{ is true}\}$ be the index set for which H_0 is true
- Define $R = \{i; H_{j0} \text{ is rejected}\}$ be the index set that we reject.
- We say that we have controlled the [familywise error rate](#) at level α if

$$P(R \cap I \neq \emptyset) = P(\text{making a false rejection}) \leq \alpha.$$

- **Bonferroni method:** for each individual test, set the level to be α/N . Let p_j be the p -value for

test H_{j0} versus H_{j1} .

$$\begin{aligned} P(\text{making a false rejection}) &= P(p_j < \alpha/N \text{ for some } i \in I) \\ &\leq \sum_{i \in I} P(p_j < \alpha/N) \\ &= \sum_{i \in I} \alpha/N = \frac{\alpha|I|}{N} \leq \alpha \end{aligned}$$

So we have overall control of the type 1 error.

- It can have low power.

Normal Example Example. Suppose we have N sample means Y_1, \dots, Y_N , each is the average of n normal observations with variance σ^2 . So $Y_j \sim N(\mu_j, \sigma^2/n)$. To test $H_{j0} : \mu_j = 0$ we can use the test statistic

$$T_j = \sqrt{n}Y_j/\sigma \sim N(\mu_j, 1).$$

The power function at $\mu_j = 0$ (p -value) is $p_j = 2\Phi(-|T_j|)$.

- If we did uncorrected testing that we reject $p_j < \alpha$, which means $|T_j| > z_{\alpha/2}$.
- With Bonferroni correction we reject when $p_j < \alpha/N$, which corresponds to

$$|T_j| > z_{\alpha/2N}$$

- Generate random samples under H_0 with code in next slide.
- If we apply the approximation for normal CDF and PDF, tha

$$\frac{\phi(x)}{x + 1/x} \leq 1 - \Phi(x) \leq \frac{\phi(x)}{x}, \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

then approximately, the corrected bound becomes $\sigma\sqrt{2\log(2N/\alpha)/n}$. It grows like $\sqrt{\log N}$.

Rcode

```

1 rm(list=ls())
2 N = 50; n = 100; sigma = 3; alpha = 0.05;
3 iter = 100; fwr1 = rep(0, iter); fwr2 = rep(0, iter);
4 for(i in 1:iter){
5   Y = rnorm(N, mean = 0, sd = sigma/sqrt(n));
6   #Generate Y_i' s under null
7   stat = sqrt(n)*Y/sigma;
8   #Calculate the test statistic T_i
9   p = 1*(abs(stat) > qnorm(1 - alpha/2))
10  #Find the p-value for each individual test without correction

```

	Decision	
	Retain H_0	Retain H_0
H_0 is true	✓ (true negative)	Type 1 error(false positive)
H_1 is true	Type 2 error(false positive)	✓ (true postive)

```

11 corp = 1*(abs(stat) > qnorm(1 - alpha/2/N))
12 #Find the p-value for each individual test with Bonferroni cor
13 fwr1[i] = 1*(sum(p) > 0); #Familywise error for test 1;
14 fwr2[i] = 1*(sum(corp) > 0); #Familywise error for test 2;
15 }
16 mean(fwr1) #empirical familywise error for uncorrected test
17 mean(fwr2) #empirical familywise error for corrected test

```

8.8.2 Familywise Error Control

Recall we have the table:

- Define the false discovery proportion as

$$FDP = \frac{|R \cup I|}{|R|} = \frac{\#FP}{\#FP + \#TP}$$

- The false discovery rate is defined as the expectation of FDP.

Our goal is to let

$$FDR = E[FDP] \leq \alpha$$

False Discovery Control Benjamin-Hochberg method:

- (1) Find the ordered p -values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(N)}$
- (2) Let $j = \max\{i : p_{(i)} < i\alpha/N\}$. Let $T = p_{(j)}$.
- (3) Let $R = \{i : p_i \leq T\}$.

Proof. For simplicity, assume we reject hypothesis j when $p_j \leq t$. Let \hat{G} be the empirical distribution of the p -values and let $G(t) = E[\hat{G}(t)]$. In this case,

$$FDP = \frac{\sum_{i=1}^N I(p_i < t, H_0)}{\sum_{i=1}^N I(p_i < t)} = \frac{\frac{1}{N} \sum_{i=1}^N I(p_i < t, H_0)}{\frac{1}{N} \sum_{i=1}^N I(p_i < t)}.$$

Hence,

$$\begin{aligned} E[FDP] &\approx \frac{E[\frac{1}{N} \sum_{i=1}^N I(p_i < t, H_0)]}{\frac{1}{N} E[\sum_{i=1}^N I(p_i < t)]} = \frac{\frac{1}{N} \sum_{i=1}^N E[I(p_i < t, H_0)]}{\frac{1}{N} \sum_{i=1}^N E[I(p_i < t)]} \\ &= \frac{t\#P/N}{G(t)} \leq \frac{t}{G(t)} \approx \frac{t}{\hat{G}(t)} \end{aligned}$$

Let $t = p_{(i)}$ for some i ; then $\hat{G}(t) = i/N$. Thus, $FDR \leq p_{(i)}N/i$. Setting it equal to α and we have the result.

8.8.3 Higher Criticism

Higher Criticism

- Let p_j be the p -value for test problem H_j . Then under null hypothesis, $p_j \sim Unif(0, 1)$.
- Consider a level α test for an individual hypothesis test, we reject the hypothesis when $p_j < \alpha$
- Let $Y_j = I(p_j \leq \alpha)$. Under null hypothesis, $Y_j \sim Bernoulli(\alpha)$. So \bar{Y}_n has mean α and standard deviation $\sqrt{\alpha(1-\alpha)/N}$

- According to CLT, let \bar{Y}_N be the fraction of the rejected hypothesis with a level α test, then

$$T_N = \sqrt{N} \frac{\bar{Y}_n - \alpha}{\sqrt{\alpha(1-\alpha)}} \xrightarrow{d} N(0, 1),$$

When $|T_n| > z_{\alpha/2}$, at least one hypothesis testing problem rejects H_0

Higher Criticism

- Why do we consider a fixed α ? When α changes, will we get different results?
- Define the function of α . The expression is as following:

(1) Sort p -values so that $p_{(1)} < p_{(2)} < \cdots < p_{(N)}$

(2) Define

$$HC_k = \sqrt{N} \frac{k/N - p_{(k)}}{\sqrt{p_{(k)}(1 - p_{(k)})}}.$$

(3) The test statistic is $T_N = \max_{1 \leq k \leq N/2} HC_k$

- Note. Now $p_{(k)}$ plays the role of α . If we take $\alpha = p_{(k)}$, then the number of rejected hypothesis is k , so the fraction is k/N .
- The limiting distribution for T_n is Gumbel distribution (no need to know)

Comparison

- The three methods cares about different things
- The Bonferoni correction is to control the familywise error, which is the exact number of Type 1 wrong decisions. It does not need independence assumption between tests, but it may lose some power.
- The FDR method cares about the fraction of false positive. It is useful when the true positives are rare.
- Higher Criticism cares about whether there is any positive or not. It works for the case that true positives are rare, and it allows the signals to be moderately weak. On the other hand, to make sure it works, the dependence between tests cannot be strong.

第 9 讲 Confidence Sets

9.1 Confidence Sets

- Related to the hypothesis testing problem, an interesting topic is the confidence sets.
- In point estimation, our estimator is $T(X_1, \dots, X_n)$
- Once we observed the data, our estimate is $T(X_1, \dots, X_n)$. It is consistent (close to the truth), yet it does not equal to the truth.
- Moreover, in most cases, $P(T(X_1, \dots, X_n) = \theta_0) = 0!$

定义 9.1 (Definition 9.1.1: Confidence Intervals)

An interval estimate for θ , is any pair of function $L : X^n \rightarrow \mathbb{R}$,

$U : X^n \rightarrow \mathbb{R}$, such that $L(x_{1:n}) \leq U(x_{1:n})$, any $x_{1:n} \in X^n$. The random interval $|L(X_{1:n}), U(X_{1:n})|$ is called an *interval estimator*.



- For an interval, we can claim the probability that it contains the true parameter.

- It is called the **coverage probability** of an interval estimator that

$$P(\theta \in [L(X_{1:n}), U(X_{1:n})]; \theta).$$

$\inf_{\theta \in \Theta} P(\theta \in [L(X_{1:n}), U(X_{1:n})]; \theta)$ is called the **confidence coefficient**.

Example Let $X_i \stackrel{i.i.d}{\sim} \text{Unif}[0, \theta], i = 1, \dots, n$. Set $Y = X_{(n)}$. We are interested in an interval estimator for θ . Consider the interval with the form $[aY, bY]$ for some $1 \leq a < b$, Then,

$$P(aY \leq \theta \leq bY; \theta) = P\left(\frac{1}{b} \leq Y/\theta \leq \frac{1}{a}; \theta\right).$$

The CDF of Y is

$$P(Y \leq c) = \left(\frac{c}{\theta}\right)^n, \quad P\left(\frac{Y}{\theta} \leq c\right) = P(Y \leq c\theta) = c^n.$$

Therefore, the coverage probability is

$$P(aY \leq \theta \leq bY; \theta) = (1/a)^n - (1/b)^n.$$

The confidence coefficient is the same.

Question: Is the confidence coefficient always the same with the coverage probability? Answer: No!

Example,2 Still consider the previous example. Now we consider the confidence interval with the form $[Y + c, Y + d]$, $0 \leq c < d$. Now the coverage probability is

$$\begin{aligned} P(Y + c \leq \theta \leq Y + d; \theta) &= P(\theta - d \leq Y \leq \theta - c; \theta) \\ &= \left(\frac{\theta - c}{\theta}\right)^n - \left(\frac{\theta - d}{\theta}\right)^n \\ &= (1 - c/\theta)^n - (1 - d/\theta)^n \end{aligned}$$

The coverage probability changes with θ . Note that

$$\lim_{\theta \rightarrow \infty} P(\theta \in [Y + c, Y + d]; \theta) = 0.$$

So the confidence coefficient is 0.

Confidence Sets General methods to get the confidence sets:

- Probability Inequalities
- Inverting a test
- Pivots

9.1.1 Probability Inequalities

Review of Probability Inequalities

- Markov Inequality: for non-negative random variable X ,

$$P(X \geq a) \leq \frac{E[X]}{a}$$

- Chebyshev's inequality. Let $\mu = E[X]$ and $\sigma^2 = \text{Var}(X)$. Then,

$$P(|X - \mu| \geq t) \leq \sigma^2/t^2$$

- Normal Tail Inequality. Let $X \sim N(0, 1)$, then we have

$$P(|X| > \epsilon) \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}$$

Proof. Set $Y = |X| \cdot 1_{\{|X| > \epsilon\}}$. Then $P(|X| > \epsilon) = P(Y > \epsilon)$.

$$E[Y] = 2 \int_{\epsilon}^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \frac{2}{\sqrt{2\pi}} (-e^{-y^2/2})|_{\epsilon}^{\infty} \leq 2e^{-\epsilon^2/2}.$$

With Markov Inequality,

$$P(|X| > \epsilon) \leq P(Y > \epsilon) \leq \frac{E[Y]}{\epsilon} < \frac{2e^{-\epsilon^2/2}}{\epsilon}.$$

Probability Inequalities

- Chernoff's inequality. Let X be a random variable. For $t \geq 0$,

$$P(|X| > \epsilon) = P(e^{tX} > e^{t\epsilon}) \leq e^{-t\epsilon} E[e^{tX}] \Rightarrow P(|X| > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon} E[e^{tX}]$$

- Hoeffding's inequality. Let X_1, \dots, X_n be i.i.d. r.v.'s with mean μ and $a \leq X_i \leq b$.

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}, \quad \epsilon > 0.$$

Example. For i.i.d. Bernoulli(p) random sample X_1, \dots, X_n , we have $E[X] = p$ and they are bounded by 0 and 1. So,

$$P(|\bar{X}_n - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

- Bernstein's Inequality. Let X_1, \dots, X_n be i.i.d. r.v.'s with mean μ , variance σ^2 and $a \leq X_i \leq b$.

Then we have

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2(\sigma^2 + (b-a)\epsilon)}} \quad \epsilon > 0.$$

For the r.v.'s that concentrate in a small interval, this bound is more helpful.

9.1.2 Inverting a Test

Confidence Intervals

- Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. By Hoeffding's inequality,

$$P(|\bar{X}_n - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

So, to construct a confidence interval with confidence coefficient $1 - \alpha$, we let $\alpha = 2e^{-2n\epsilon^2}$, and solve it with $\sqrt{\log(2/\alpha)/2n}$. For the interval $[\bar{X} - \epsilon, \bar{X} + \epsilon]$, we have

$$\begin{aligned} P(\bar{X} - \epsilon \leq p \leq \bar{X} + \epsilon; p) &= P(-\epsilon \leq \bar{X} - p \leq \epsilon) \\ &= P(|\bar{X} - p| \leq \epsilon) \geq 1 - 2e^{-2n\epsilon^2} = 1 - \alpha. \end{aligned}$$

- Now, consider the Poisson distribution. Let $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$. We want to construct a confidence interval for λ .

Recall that $\sum X_i \sim \text{Poisson}(n\lambda)$, with mean $n\lambda$ and variance $n\lambda$. With Chebyshev's inequality, there is

$$P(|\bar{X}_n - \lambda| \geq \epsilon) \leq \lambda/n\epsilon^2.$$

Set $\alpha = \lambda/n\epsilon^2$, which solves that $\epsilon_n = \sqrt{\lambda/n\alpha}$. The $1 - \alpha$ -confidence intervals is $[\bar{X} -$

$$\sqrt{\lambda/n\alpha}, \bar{X} + \sqrt{\lambda/n\alpha}].$$

Inverting a Test

- Consider the Hypothesis testing problem

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Say that we have a test statistic T and rejection region R . We consider level α test, so that $P(T \in R; \theta_0) \leq \alpha$, and so $P(T \notin R; \theta_0) \geq 1 - \alpha$.

- Define the acceptance region $A(\theta_0)$, where $A(\theta_0)$ is the set in X^n .

$$A(\theta_0) = \{(x_1, \dots, x_n) : T((x_1, \dots, x_n)) \notin R(\theta_0)\}.$$

- Define the confidence set. The confidence set is a set in the parameter space Θ , defined by the observations (x_1, \dots, x_n) .

$$C_n = C_n(x_1, \dots, x_n) = \{\theta : (x_1, \dots, x_n) \in A(\theta)\}.$$

- Coverage Probability:

$$\begin{aligned} P(\theta \in C; \theta) &= P((X_1, \dots, X_n) \in A(\theta); \theta) \\ &= P(T(X_1, \dots, X_n) \notin R(\theta); \theta) \geq 1 - \alpha. \end{aligned}$$

Inverting a Test

- The procedure seems hard to understand, yet the procedure is easy
- Let $X_1, \dots, X_n \sim N(\theta, 1)$, The LRT of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ has rejection region as

$$|\bar{X} - \theta_0| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

So, the acceptance region is $A(\theta)$ is a set about $x_{1:n}$, which changes with θ

$$A(\theta) = \{(x_1, \dots, x_n); |\bar{x} - \theta| < \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\},$$

and so $\theta \in C(X^n)$ if and only if

$$|\bar{X} - \theta_0| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

In other words, the confidence interval is $(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2})$. This interval has confidence coefficient as $1 - \alpha$.

Inverting a Test

- As long as we have a test, we can find the confidence interval with it. It applies for whatever test, the Wald test, the Neyman-Pearson test, the t test and the F-test, etc.

- With this procedure, it is possible that we cannot get an interval. That's why we call it "confidence sets" instead of confidence intervals
- With a $1 - \alpha$ confidence set $C(x_1, \dots, x_n)$, we can also figure out a test:

$$\text{reject } H_0 : \theta = \theta_0 \text{ if } \theta_0 \notin C(x_1, \dots, x_n).$$

It is a level α test.

- However, it is much less used. The most general direction is from hypothesis testing problems to the confidence interval estimation, i.e., the distribution is the same for every $\theta \in \Theta$.

9.1.3 Pivot

Pivot

Definition: Pivot

A function $Q(X_1, \dots, X_n, \theta)$ is a pivot if the distribution of Q does not depend on θ .

- If the distribution of Q is known, with the relationship between X_1, \dots, X_n and θ in Q , we can

build a confidence interval.

- Let a and b be such that

$$P(a \leq Q(X, \theta) \leq b) \geq 1 - \alpha.$$

The confidence interval follows as $C(x) = \{\theta : a \leq Q(X, \theta) \leq b\}$

- Example. $N(0, 1)$ distribution. $\bar{X} - \theta \sim N(0, 1/n)$, which does not depend on θ
- Any location families has pivot as $\bar{X} - \theta$.

Example Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Unif}(0, \theta)$. Let $Q = X_{(n)}/\theta$. Then the CDF of Q is

$$P(Q \leq t) = \prod_{i=1}^n P(X_i \leq t\theta) = \left(\frac{t\theta}{\theta}\right)^n = t^n, \quad 0 < t \leq 1.$$

It does not depend on θ , so Q is a pivot.

To find a $1 - \alpha$ confidence interval, note that

$$P(c \leq Q \leq 1) = 1 - P(Q \leq c) = 1 - c^n.$$

Let $1 - \alpha = 1 - c^n$, then $c = \alpha^{1/n}$.

$$P(c \leq Q \leq 1) = 1 - c^n = 1 - \alpha.$$

The $1 - \alpha$ confidence interval is

$$\begin{aligned} C(X_{1:n}) &= \{\alpha^{1/n} \leq X_{(n)}/\theta \leq 1\} = \{X_{(n)} \leq \theta \leq X_{(n)}/\alpha^{1/n}\} \\ &= (X_{(n)}, X_{(n)}/\alpha^{1/n}) \end{aligned}$$

9.1.4 Confidence Sets of CDF

~~Confidence Sets of CDF~~ Let $X_1, \dots, X_n \sim F$. In Lecture 1, we referred the empirical CDF,

$$\hat{F}(x) = \frac{1}{n} \sum 1\{X_i \leq x\}.$$

This is an estimation. Can we find the confidence sets for $F(x)$?

- This is nonparametric estimation. Yet we can still apply the parametric approximations
- For fixed x , note that $\hat{F}(x)$ is the average of n Bernoulli($F(x)$), we can apply the confidence interval results for the Bernoulli random variables
- We are interested in the confidence sets for the whole CDF. We want to figure out $L(x)$ and

$U(x)$, so that

$$P(L(x) \leq F(x) \leq U(x) \text{ for all } x) \geq 1 - \alpha$$

Pivot Empirical CDF: $\hat{F}(x) = \frac{1}{n} \sum 1\{X_i \leq x\}$.

- Let $K_n = \sup_x |\hat{F}(x) - F(x)|$. K_n measures the largest difference between the empirical CDF and the truth. Once K_n is properly bounded, the confidence sets for $F(x)$ among all x can be fixed.
- For continuous F , K_n is a pivot. To see this, let $U_i = F(X_i)$. Then $U_1, \dots, U_n \stackrel{i.i.d}{\sim} \text{Unif}(0, 1)$. So,

$$\begin{aligned} K_n &= \sup_x |\hat{F}(x) - F(x)| = \sup_x \left| \frac{1}{n} \sum 1\{X_i \leq x\} - F(x) \right| \\ &= \sup_x \left| \frac{1}{n} \sum 1\{F(X_i) \leq F(x)\} - F(x) \right| \\ &= \sup_x \left| \frac{1}{n} \sum 1\{U_i \leq F(x)\} - F(x) \right| \\ &= \sup_{0 \leq t \leq 1} \left| \frac{1}{n} \sum 1\{U_i \leq t\} - t \right| \end{aligned}$$

The result does not depend on F .

- Find a number c , so that $P(\sup_{0 \leq t \leq 1} |\frac{1}{n} \sum 1\{U_i \leq t\}| > c) = \alpha$.
- The confidence set is then $C = \{F : \sup_x |F_n(x) - F(x)| < c\}$.

Credible Sets In Bayesian statistics, what is the confidence set?

- Recall. For Bayesian statistics, the parameters are not constants. There is a prior $\pi(\theta)$ for the parameter θ
- With the observed data, we update the prior $\pi(\theta)$ to the posterior $\pi(\theta|X)$
- If we have a loss function, we summarize $\pi(\theta|X)$ into an estimator with smallest Bayes risk.
- However, for Bayesian statisticians, $\pi(\theta|X)$ is the estimation for the parameter θ
- Confidence sets: the probability that the estimated set include the true parameter θ_0
- In Bayesian, there is no *truth*. They update the prior distribution with more and more data, to get a more and more accurate posterior distribution. So, no *confidence interval* thing!
- Yet, there is so-called *credible sets*

Credible Sets

- Assume we observe a random sample $X_1, \dots, X_n \sim F(x; \theta)$, and the prior is $\pi(\theta)$
- With the data, we have the posterior $\pi(\theta)$

- The $1 - \alpha$ [credible set](#) C is defined as

$$P(L(X_{1:n}) \leq \theta \leq U(X_{1:n}) | X) \geq 1 - \alpha.$$

- We still have a set here. The set has probability $1 - \alpha$
- Difference: For confidence set, θ is fixed, $L(X)$ and $U(X)$ are random. The probability is the probability that (L, U) contains θ . If we draw the samples again and again, then the probability it covers θ is $1 - \alpha$. For credible sets, θ is random. With the given data, we are interested in the interval that θ concentrates on.
- To find the credible set, just figure out the posterior distribution, and draw an interval for θ with probability $1 - \alpha$.